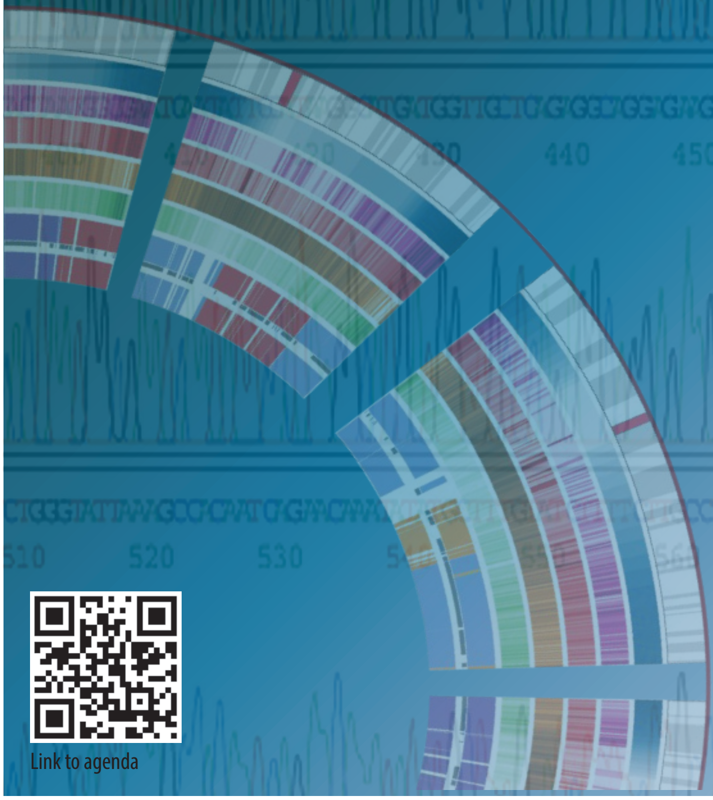
10th Annual Sequencing, Finishing, Analysis in the Future Meeting

Santa Fe, New Mexico
May 27-29, 2015



Link to agenda

Los Alamos
NATIONAL LABORATORY
EST. 1943

Table of Contents

The 2015 "Sequencing, Finishing, and Analysis in the Future" Organizing Committee

* Chris Detter, Ph.D., Senior Science Advisor, DTRA

* Johar Ali, Ph.D., Research Director, International, Alviarmani

* Patrick Chain, Bioinformatics/Metagenomics Team Leader, LANL

* Michael Fitzgerald, Microbial Special Projects Manager, Broad Institute

* Bob Fulton, M.S., Director of Project Development & Management, WashU

* Darren Grafham, Lab Manager, Children's Hospital, Sheffield, UK

* Alla Lapidus, Ph.D., Director, Genomics, Algorithmic Biology Lab, SPbAU; Russia

* Donna Muzny, M.Sc., Director of Operations, Baylor College of Medicine

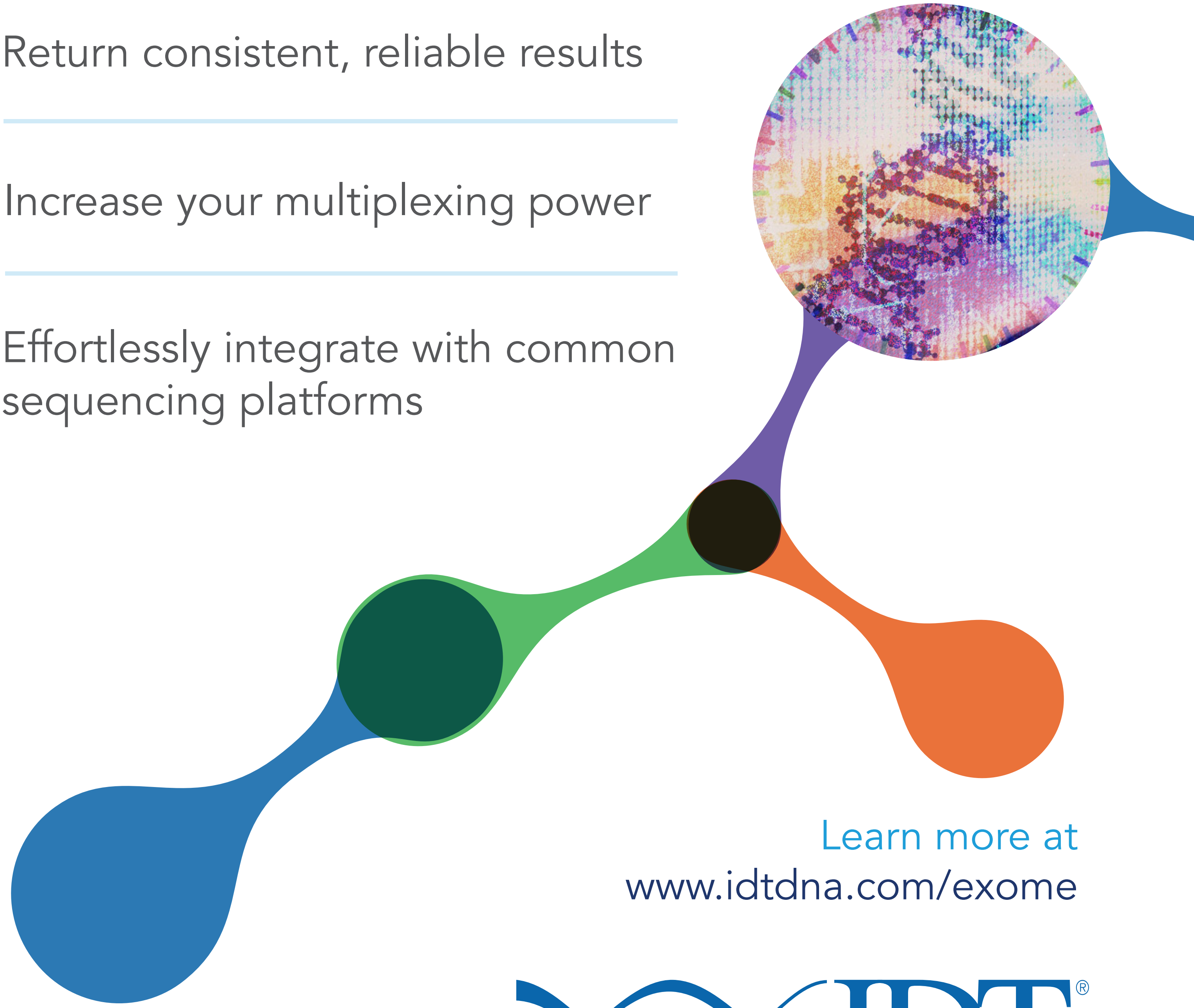
* Shannon Dugan-Perez, Project Manager, Baylor College of Medicine

* David Bruce, Project and Program Manager of Genomic Sciences, LANL

* Shannon Johnson, Ph.D., Project Manager, LANL

10th Annual SFAF Agenda Overview

Wednesday, 27th May		
07:30 - 08:30	*American Breakfast Buffet*	*Sponsor: NEB*
08:30 - 08:45	Welcome Introduction from Los Alamos National Laboratory	TBD
08:45 - 09:30	Keynote Address: The Continuous Spectrum of Mutations Spanning Pediatric & Adult Disease (Dr. Eric Boerwinkle)	Sponsor: LabCyte
09:30 - 10:10	**Oral Session 1: NGS Applications & Analysis (Chair: Donna Muzny & Johar Ali)**	
10:10 - 10:40	*Coffee Break*	*Sponsor: Promega*
10:40 - 11:25	**Tech Session 1 Sequence Platform and Performance Updates** **(Chair: Bob Fulton & Patrick Chain)**	
11:25 - 12:00	**NGS 10yr Flashback and Look-forward** **Round Table 1, Panel: Haley Fiske, Abizar Lakdawalla, Steve Turner and Tim Harkins** **(Chair: Bob Fulton & Patrick Chain)**	
12:00 - 13:30	*Coronado Lunch Buffet*	*Sponsor: PacBio*
13:30 - 15:30	**Oral Session 2 NGS Applications & Analysis (Chair: Alla Lapidus & Darren Grafham)**	
15:30 - 16:00	*Coffee Break*	*Sponsor: PerkinElmer*
15:30 - 18:00	**Tech Talks: Sequencing Technologies (Chair: Darren Grafham & Alla Lapidus)**	
18:30 - 21:30	**Poster Sessions with Meet & Greet**	Sponsor: Roche
Thursday, 28th May		
07:30 - 08:30	*La Fonda Breakfast Buffet*	*Sponsor: NEB*
08:30 - 08:45	Welcome Introduction & Opening Remarks	
08:45 - 09:30	Keynote Address: In Search of the Perfect Assembly (Dr. Daniel Rokhsar)	Sponsor: ThermoFisher
09:30 - 10:30	**Oral Session 3 Genome Assembly & Analysis (Chair: Alla Lapidus & Patrick Chain)**	
10:30 - 11:00	*Coffee Break*	*Sponsor: BioNano Genomics*
11:00 - 12:20	**Oral Session 4 Genome Assembly & Analysis (Chair: Mike Fitzgerald & Bob Fulton)**	
12:20 - 13:50	*New Mexican Lunch Buffet*	*Sponsor: Promega*
13:50 - 15:30	**Oral Session 5 Genome Assembly & Analysis (Chair: Darren Grafham & Donna Muzny)**	
15:30 - 15:45	*Coffee Break*	*Sponsor: Lucigen*
15:45 - 17:30	**Tech Talks: Assembly & Analysis (Chair: Johar Ali & Mike Fitzgerald)**	
18:00 - 20:00	*Happy Hour Cowgirl Cafe*	*Sponsor: Illumina*
Friday, 29th May		
07:30 - 08:30	*Harvey House Breakfast*	*Sponsor: NEB*
08:30 - 08:45	Opening Remarks	
08:45 - 09:30	Keynote Address: Evolution and Epidemiology of Anthrax through lens of Genome Analysis (Dr. Paul Keim)	Sponsor: Advanced Analytic
09:30 - 10:30	**Oral Session 6 Pathogens & Microbial Genomics (Chair: Donna Muzny & Bob Fulton)**	
10:30 - 10:45	*Coffee Break*	*Sponsor: PacBio*
10:45 - 12:45	**Oral Session 7 Pathogens & Microbial Genomics** **(Chair: Mike Fitzgerald & Patrick Chain)**	
12:45 - 13:30	*Santa Fe Deli Lunch Buffet*	*Sponsor: MRIGlobal*
12:45 - 13:00	CR-1 Closing Remarks (End of Regular Sessions) Wrap up of SFAF2015 and planning for 2016	Chris Detter
13:30 - 15:00	**Forensic Analysis 1 Forensic Applications of NGS** **(Chair: Robert Bull & Kristen McCabe)**	
15:00 - 15:15	*Coffee Break*	*Sponsor: Qiagen*
15:15 - 17:00	**Forensic Analysis 2 Forensic Applications of NGS** **(Chair: Robert Bull & Kristen McCabe)**	
16:45 - 17:30	**Round Table 2: Discussion of Forensics Applications for NGS Technologies** **(Chair: Robert Bull)**	



Wednesday, May 27th Agenda

07:30 - 08:30	*American Breakfast Buffet*	*Sponsor: NEB*
08:30 - 08:45	Welcome Introduction from Los Alamos National Laboratory	TBD
08:45 - 09:30	Keynote Address: The Continuous Spectrum of Mutations Spanning Pediatric and Adult Disease (Dr. Eric Boerwinkle)	Sponsor: LabCyte
09:30 - 10:10	**Oral Session 1: NGS Applications & Analysis** **(Chair: Donna Muzny & Johar Ali)**	
09:30 – 09:50	Applications of Next Generation Sequencing (NGS) in Newborn Screening (NBS) in the UK National Health Service (NHS)	Darren Grafham
09:50 – 10:10	Preparing (for) Cohorts of X Ten Whole Genomes	Will Salerno
10:10 - 10:40	*Coffee Break*	*Sponsor: Promega*
10:40 - 11:25	**Tech Session 1 Sequence Platform and Performance Updates** **(Chair: Bob Fulton & Patrick Chain)**	
10:40 – 10:55	Illumina Update on Sequencing and Performance	Kelly Hoon
10:55 – 11:10	Ion Torrent Semiconductor Sequencing Performance Update	Mike Lelivelt
11:10 – 11:25	PacBio Platform and Performance Update	Steve Turner
11:25 - 12:00	**NGS 10yr Flashback and Look-forward: Round Table 1** **Panel: Haley Fiske, Abizar Lakdawalla, Steve Turner and Tim Harkins** **(Chair: Bob Fulton & Patrick Chain)**	
12:00 - 13:30	*Coronado Lunch Buffet*	*Sponsor: PacBio*
13:30 - 15:30	**Oral Session 2 NGS Applications & Analysis** **(Chair: Alla Lapidus & Darren Grafham)**	
13:30 – 10:50	Interactive analysis and quality assessment of single-cell copy number variations	Tyler Garvin
13:50 – 14:10	100K Pathogen Genomes Project: Progress, Expansion, and Interrogation	Dylan Storey
14:10 – 14:30	Whole Genome Focused Array SNP Typing: A method for characterizing Mycobacterium tuberculosis molecular epidemiology from direct clinical samples using whole genome sequencing	Rebecca Colman
14:30 – 14:0	Diagnosis of Acute Respiratory Viral Infections by Targeted RNA Sequencing Provides Additional Critical Genetic Virulence and Epidemiologic Information	Darrell Dinwiddie
14:50 – 15:30	Recovery of Thauera sp. SWB20 draft genome isolated from a Singapore wastewater treatment facility using gel microdroplets and single cell genomics	Armand Dichosa
15:30 - 16:00	*Coffee Break*	*Sponsor: PerkinElmer*
15:30 - 18:00	**Tech Talks: Sequencing Technologies** **(Chair: Darren Grafham & Alla Lapidus)**	
15:30 – 15:45	Long Range Applications from Short Read Sequencing	Rob Tarbox
15:45 – 16:00	Automatic Closing and Finishing of Genomes with Long Mate Pair NGS Libraries	David Mead
16:00 – 16:15	Reducing Bias in Small RNA-Sequencing	Masoud Toloue
16:15 – 16:30	A novel in vitro method for obtaining high-quality, long-range genomic information	Brandon Rice
16:30 – 16:45	A novel, streamlined NGS library preparation workflow employing enzymatic fragmentation results in significant improvements to library yields and sequence quality	Jennifer Pavlica
16:45 – 17:00	Oligonucleotides for Hybridization-Based Target Enrichment	John Havens
17:00 – 17:15	AmpliSeq: targeted, high sensitivity sequencing for tumor mutation detection at 0.5%	Yongming Sun
17:15 – 17:30	A technology platform for microbial genomics	Sam Minot
17:30 – 17:45	Structural Variant Detection with Single Molecule Solid-State Detectors	John Oliver
18:30 - 21:30	**Poster Sessions with Meet & Greet**	Sponsor: Roche
18:30 - 20:00	Poster Sessions -1a & 2a (1a on Poster Session 1st Floor & 2a on 2nd Floor)	
20:00 - 21:30	Poster Sessions -1b & 2b (1b on Poster Session 1st Floor & 2b on 2nd Floor)	

The Continuous Spectrum of Mutations Spanning Pediatric and Adult Disease

Wednesday, 27th May 8.45 - La Fonda Ballroom (Sponsored by LabCyte) - Keynote/Plenary

Eric Boerwinkle
University of Texas, Health Science Center at Houston

Human geneticists have traditionally created and worked within defined domains, such as pediatric and adult genetics or Mendelian and complex disease genetics. However, a contemporary view emerging as a result of exome or whole genome sequencing and analysis of large numbers of clinically defined patients or deeply phenotyping cohort studies shows that these bins are convenient constructs but do not reflect the reality of the genetic architecture of health and disease. The Baylor College of Medicine CLIA/CAP sequencing laboratory (Whole Genome Laboratory: WGL), as of 01/31/2015, has analyzed more than 5,000 cases of childhood disease by exome sequencing. Approximately 11% of the WGL patients are adults (>18 yrs), most often referred by clinical geneticists and clinically representing neurologic conditions. We report an ~25% solution rate, including multiple instances of rare and de novo events. We confirm a ~5% prevalence of clinically ascertained diagnostic dilemmas having blended phenotypes confounded by mutations at multiple loci. Genomic approaches revealed such individuals to be harboring variants at two genetic loci that otherwise would each be predicted to cause disease. Appropriately consented unsolved cases create a virtual cycle of novel gene discovery, which ultimately improves the overall diagnostic rate, as well as development of new tools, such as computational methods to identify pathogenic single-exon deletions, that can be progressively adapted for clinical application. Approximately 30% of diagnosed cases were in genes discovered since 2011. In the Center for Mendelian Genomics we have sequenced ~4,500 individuals leading to ~320 discoveries, including novel genes and novel genotype-phenotype relationships (i.e. phenotype expansion). As sequencing becomes a common tool in the health care setting, we have analyzed the exomes of ~8,500 individuals belonging to deeply phenotyped cohort studies. Loss-of-function mutations leading to extreme phenotypes have led to multiple novel gene discoveries (e.g. HAL and cardiovascular disease), and is an ideal platform for Mendelian randomization experiments of potential drug targets (e.g. PLA2G7). Regular clinical application of genome sequencing is emerging as a strong platform for novel discovery and so the complete model represents both the demonstration of a full translation of genomics into the clinic, and integration and merging of the research and clinical goals.

Author credit: Eric Boerwinkle1,2, Yaping Yang3, Christine M Eng3, Donna M Muzny2, James R Lupski2,3,4, Jennifer E Posey3, Alanna Morrison1, Richard A Gibbs2
1Human Genetics Center, University of Texas, Houston, TX; 2Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX; 3Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX; 4Department of Pediatrics, Baylor College of Medicine, Houston, TX

Applications of Next Generation Sequencing (NGS) in Newborn Screening (NBS) in the UK National Health Service (NHS)

Wednesday, 27th May 9.30 - La Fonda Ballroom - Oral

Darren Grafham[1]
[1]***Sheffield Diagnostic Genetics Service on behalf of Next Generation Sequencing in Newborn Screening Project***

Newborn screening (NBS) programmes are utilised worldwide to identify babies affected with rare, often fatal, but treatable disorders. Biochemical analysis of dried blood spot (DBS) samples provides cost-effective testing in a short turnaround time. Screen-positive infants receive rapid medical attention, but predicting disease severity and thus appropriate management in a currently asymptomatic individual can be challenging. Current NBS precludes analysis of disorders lacking a suitable biochemical analyte or enzyme activity assay.

Sheffield Diagnostic Genetics Service is the most integrated NHS diagnostic genetics department in the UK and is investigating the use of Next-Generation Sequencing (NGS) to: 1) improve the diagnostic and prognostic utility of existing UK NBS programmes, and 2) assess the technical feasibility of using NGS as the primary screening modality. Enhanced diagnosis and prognosis for existing programmes would facilitate personalised treatment and prevention of unnecessary medicalisation of screen-positive, clinically unaffected children.

Preparing (for) Cohorts of X Ten Whole Genomes

Wednesday, 27th May 9.50 - La Fonda Ballroom - Oral

Will Salerno[1], Adam English[1], Matthew Bainbridge[1], Mike Dahdouli[1], Simon White[1], Xiaoming Liu[2], Narayanan Veeraraghavan[1], Donna Muzny[3], Eric Boerwinkle[3], Richard Gibbs[3]

[1]Human Genome Sequencing Center, Baylor College of Medicine, [2]Human Genetics Center, University of Texas Health Science Center, [3]Baylor College of Medicine

To improve high-throughput whole-genome sequencing analysis we developed Mercury II, a whole-genome, petabase-scale upgrade to our Mercury infrastructure that calls and prioritizes small and structural variants using consensus methods, whole-genome annotation sources, and comparison to large control cohorts and robustly characterized personal genomes. Mercury II comprehensive genomes provide the appropriate biological, epidemiologic and medical contexts, with special attention to previously under-analyzed genomic features such as non-coding regions and large, complex variants.

Mercury II is optimized for the petabases of whole-genome data generated by the Illumina HiSeq X platform. Mapping, quality control, and variant calling have a raw-data-to-annotated-variants turnaround time of less than a week. Mercury II SNVs and indels are called with ATLAS3 and annotated with whole-genome genomic features including regulatory features from ENCODE, GTEx, promoter and enhancer sites from FANTOM5 and deleteriousness scores from CADD, RegulomeDB, and Funseq2. Structural variants (SVs) are identified with Parliament, a consensus SV tool that automates assembly-based force calling. Mercury II re-analysis options include unaligned read remapping to non-reference sources, integrated RNA-seq and miRNA-seq, teleomere assessment, and priority region refinement (read-stitching, local assembly, deep SV calling). Variants are stored in a data warehouse containing results from a high-confidence gold genome and large cohorts that can be used for comparison to ethnic-matched case groups. All variants are prioritized based on annotation and commonality to large cohorts and the gold genome, which is especially relevant to SV analysis given the relative dearth of available annotation compared to that of smaller variants.

To validate Mercury II, we analyzed 10 deep whole-genome trios. ATLAS3 and Parliament identified ~500 putative de novo events per family. Of these, less than 10 per variant type (SNV, indel, SV) intersected protein-coding regions. Annotation summaries of all variants are also provided. Finally, these findings are placed in the context of similar findings within the ARIC cohort and Mercury II titrations of the HS1011 genomic data, which include 100 bp and 150 bp (PCR-free) short reads and ~10 kbp long reads.

Mercury II delivers to users "BAM-free" project-ready summary files, visualization tools, and an automated re-analysis toolbox.

Illumina Update on Sequencing and Performance

Wednesday, 27th May 10.40 - La Fonda Ballroom - Tech Talk

***Kelly Hoon**[1]*

*[1]**Illumina, Inc***

Ion Torrent Semiconductor Sequencing Performance Update

Wednesday, 27th May 10.55 - La Fonda Ballroom - Tech Talk

***Mike Lelivelt*[1]**

[1]*Thermo Fisher Scientific*

Abstract – Ion Torrent invented the first device—a new semiconductor chip—capable of directly translating chemical signals into digital information. The Ion Personal Genome Machine™ Sequencer, launched in December of 2010, defined the benchtop sequencer category. Across the brief 4+ years of the commercial existence of semiconductor sequencing, many advancements have been introduced including workflow enhancements with One Touch 2 and Ion Chef. Recently new enzymology has been introduced that enable increases in the fundamental accuracy of the system. The HiQ enzyme is now commercially available for both the PGM and Proton Sequencers and results in drastic decreases in false positive rates. This presentation will review data illustrating these comparison as well as accuracy performance comparison against other competitive benchtop sequencers. Ion Torrent has also continued to make large investments in improving the user experience associated with the informatics associated with Next Generation Sequencing. A review of current improvements in Ion Reporter Software and the Thermo Fisher Cloud platform will be presented.

PacBio Platform and Performance Update

Wednesday, 27th May 11.10 - La Fonda Ballroom - Tech Talk

Steve Turner[1]
[1]Pacific Biosciences

10 years of SFAF, a look at our Past and pondering our Future

Wednesday, 27th May 11.25 - La Fonda Ballroom - Panel

Haley Fiske[1], Abizar Lakdawalla[2], Steve Turner[3], Tim Harkins[4]
[1]Illumina, Inc, [2]Thermo Fisher Scientific, [3]Pacific Biosciences, [4]Swift BioSciences

A lot has happended in the NGS field since we started this meeting 10 years ago. This panel will flashback over the past 10yrs of Next Gen Sequencing and look forward to where we might be in another 10 years.

This discussion is meant to be inclusive to all those attending, so please bring lots of questions and comments to the session. The harder, more thought provoking, and more provokative the better. We (the organzing committee) want to see the panel members squirm! ☺

Interactive analysis and quality assessment of single-cell copy number variations

Wednesday, 27th May 13.30 - La Fonda Ballroom - Oral

***Tyler Garvin**[1], **Robert Aboukhalil**[1], **Michael Schatz**[1]*
*[1]**Cold Spring Harbor Laboratory***

Single-cell sequencing is emerging as a critical technology for understanding the biology of cancer, neurons, and other complex systeHere we introduce Ginkgo, a web platform for the interactive analysis and quality assessment of single-cell copy-number alterations. Ginkgo fully automates the process of binning, normalizing, and segmenting mapped reads to infer copy number profiles of individual cells, as well as constructing phylogenetic trees of how those cells are related. We validate Ginkgo by reproducing the results of five major single-cell studies, and discuss how it addresses the wide array of biases that affect single-cell analysis. We also examine the data characteristics of three commonly used single-cell amplification techniques: MDA, MALBAC, and DOP-PCR/WGA4 through comparative analysis of 9 different single-cell datasets. We conclude that DOP-PCR and MALBAC provide the most uniform amplification, while MDA introduces substantial biases into the analysis. Furthermore, given the same level of coverage, our results indicate that data prepared using DOP-PCR can reliably call CNVs at higher resolution than data prepared using MALBAC. Ginkgo is freely available at http://qb.cshl.edu/ginkgo.

100K Pathogen Genomes Project: Progress, Expansion, and Interrogation

Wednesday, 27th May 13.50 - La Fonda Ballroom - Oral

***Dylan Storey*[1], Carol Huang[1], Whitney Ng[1], Kao Thao[1], Nguyet Kong[1], Alli Weis[1], Poyin Chen[1], Narine Arabyan[1], Richard Jeannotte[1], Soraya Foutouhi[1], Azarene Foutouhi[1], Louis Sorieul[1], Bart Weimer[1]**

[1]university of california, Davis

The 100 Pathogen Project is a public/private/industry consortium addressing food safety concerns through collaborative large-scale sequencing projects aimed at public health and the food supply. This initiative's goal is to develop a high quality and publicly available genomic resource for systems biology and population scale microbiology using advanced bio-informatic approaches. To date 4233 samples have been sequenced and placed in the SRA, with approximately another 4000 submitted for sequencing. This includes 25 closed genomes generated through PacBio sequencing. We are poised to sequence an additional 5000 biological samples to be processed and sequenced in the next year at UC Davis. With the global nature of microbiology the project has expanded to China with a collaboration with Beijing Normal University that will sequence 10,000 genomes in China.

The generation and curation of these types of large datasets is fraught with technical hurdles. As a result we have developed a number of reproducible, high throughput, lab work-flows for: bacterial lysis, nucleic acid extraction, library preparation, and associated quality control metrics. By carefully developing and employing these methods with robotic automation we are able to produce 600 high-quality libraries daily. Our success at generating high quality sequencing libraries in high throughput fashion has allowed us to expand into collaborations with a number of international initiatives with similar goals in mind.

These sequencing efforts combined with publicly available datasets has necessitated updated methods for tracking the current state of sequencing in specific bacterial groups, providing diagnostic information on the resulting information from sequencing projects, and providing novel methods for interrogating many thousands of genomes in biologically meaningful ways with limited computational platforA key question we are attempting to answer is: how many genomes does it take to capture a bacterial group's complete genomic diversity. We have developed computational methods that address this question using a population scale approach. Additionally, directed analyses are possible to answer biological questions in a comprehensive genomic scale that impact public health.

Whole Genome Focused Array SNP Typing: A method for characterizing Mycobacterium tuberculosis molecular epidemiology from direct clinical samples using whole genome sequencing

Wednesday, 27th May 14.10 - La Fonda Ballroom - Oral

***Rebecca Colman*[1], *Jason Sahl*[1], *Nathan Hicks*[1], *Chandler Roe*[1], *Julia Anderson*[1], *Donald Catanzaro*[2], *Ted Cohen*[3], *Valeriu Crudu*[4], *Timothy Rodwell*[5], *Antonino Catanzaro*[5], *Paul Keim*[6], *David Engelthaler*[1]**

[1]Translational Genomics Research Institute, [2]University of Arkansas College of Education and Health Professions, [3]Yale School of Public Health, [4]Phthisiopneumology Institute (PPI), [5]University of California San

The field of molecular epidemiology has rapidly advanced with whole genome sequencing (WGS) of bacterial DNA using next-generation platforHowever, applying WGS directly to environmental or clinical samples has lagged due to the complexity of biological samples and low bacterial loads. Tuberculosis (TB) is an important infectious disease in both developed and developing countries. Understanding Mycobacterium tuberculosis (Mtb) transmission dynamics, and quickly identifying the source of an infection in an outbreak is critical to infection control, but is difficult to achieve with WGS methods based on TB cultures due to the slow growing nature of Mtb. Currently one major hurdle in the utilization of WGS for direct sample analysis is the bioinformatics resources needed for appropriate analysis of the data. We have developed an in-house pipeline, “Whole Genome Focused Array SNP Typing” (WG-FAST), for strain typing Mtb rapidly and directly from a sputum sample, which could significantly accelerate and improve existing contact tracing efforts. While a large percent of the sequence data obtained from a sputum sample may consist of human DNA reads, the use of our bioinformatic approach allows for the placement of existing phylogenetic signal on a well-characterized Mtb tree and provides confidence levels for the placement. To demonstrate the potential of our method we employed metagenome sequencing on DNA extracted from limited patient sputa from Moldova. While the percent genome coverage varied between samples, we were still able to identify sufficient SNPs using this approach, allowing for confident placement of samples on an initial tree, to identify clonal membership and/or nearest neighbors. This analysis demonstrates the feasibility of the assay despite having DNA of varying concentration and quality from direct clinical samples. Using rapid DNA sequencing with automated phylogenetic analysis for strain typing Mtb isolates from direct clinical samples allows for rapid clinical strain typing and molecular epidemiological analysis.

Diagnosis of Acute Respiratory Viral Infections by Targeted RNA Sequencing Provides Additional Critical Genetic Virulence and Epidemiologic Information

Wednesday, 27th May 14.30 - La Fonda Ballroom - Oral

Walter Dehority[1], Kimberly Paffett[1], Kevin Harrod[2], Steve Gross[3], Gary Schroth[3], Steve Young[4], Darrell Dinwiddie[1]

[1]University of New Mexico Health Sciences Center, [2]University of Alabama Birmingham School of Medicine, [3]Illumina, Inc, [4]TriCore Reference Labs

Respiratory infections cause the greatest morbidity and mortality of all pediatric ailments. Despite the enormous medical burden caused by respiratory viruses the specific genetic variation that influence transmission, virulence, and pathogenesis are poorly understood for most viruses. The objective of this study is to develop a clinical test that can both detect infections and identify genetic variation that may influence these critical processes while providing epidemiologic surveillance information. We have developed the UNM ResVir Panel, a novel, hybridization-based method to target and enrich for complete coding sequences from 34 respiratory viruses directly from clinical nasopharyngeal samples. The targets are then sequenced in a rapid manner on the Illumina MiSeq. A custom bioinformatic pipeline is used to determine the specific virus, construct nearly complete genome sequences, detect genetic variation, and identify the closest related, sequenced virus. This method and analysis provides unprecedented genetic and epidemiologic information for respiratory viral pathogens. We have used this method to identify 134 viral infections representing 12 different viruses from de-identified, residual nasopharyngeal swabs obtained from TriCore Reference Laboratories. Genomic variant characterization and whole genome phylogenetic analysis by completing hierarchical clustering to previously sequenced viral strains has revealed the presence of novel viral strains, numerous co-circulating viral strains during the same respiratory virus season, and genetic mutations that may influence pathogenesis. Our results suggest that RNA/DNA enrichment and next-generation sequencing can be used for the detection and characterization of acute respiratory infections in a culture independent, high-throughput, low cost manner. Furthermore, the method can; identify specific viral strain(s), detect genetic resistance to therapeutics, detect genetic mutations that may influence virulence, and will provide epidemiologic information.

Recovery of Thauera sp. SWB20 draft genome isolated from a Singapore wastewater treatment facility using gel microdroplets and single cell genomics

Wednesday, 27th May 14.50 - La Fonda Ballroom - Oral

Armand Dichosa[1], Karen Davenport[1], Po-E Li[1], Sanaa Ahmed[1], Hajnalka Daligault[1], Cheryl Gleasner[1], Yuliya Kunde[1], Kim McMurry[1], Chien-Chi Lo[1], Krista Reitenga[2], Ashlynn Daughton[1], Xiaohong Shen[3], Seth Frietze[4], Dongping Wang[1], Shannon Johnson[1], Daniela Drautz-Moses[5], Stephan Schuster[5], Patrick Chain[1], Cliff Han[1]

[1]Los Alamos National Laboratory, [2]Univy of Maryland School of Medicine, [3]Beacon Analytical System, Inc., [4]Univ. of Northern Colorado, [5]Nanyang Technological Univ.

Most environmental bacteria are recalcitrant to growth under traditional cultivation techniques, likely due to unknown factors such as cell growth signals or metabolites lacking in artificial media. As obtaining sufficient genomic template greatly improves bacterial genome assemblies, alternative cultivation methods are warranted to overcome the growth barrier. Here, we describe the use of gel microdroplets (GMDs) with single cell genomics (SCG) for co-cultivation of a bacterial consortium inhabiting a wastewater treatment system, resulting in the 4.93 Mbp draft genome recovery of a novel Thauera strain.

We adapted our in vitro GMD/SCG protocol to recover as much diverse bacterial representatives from the complex, sewage-wastewater community as possible. We single-captured bacteria in GMD and co-cultivated in 5% of the native wastewater and in 10% R2A broth medium. After incubation, we flow sorted 176 GMDs, each containing a single clonal microcolony, from each culture condition. Whole genome amplification and subsequent 16S rRNA phylotyping recovered 33 bacterial genera from the “native” medium and 13 bacterial genera from R2A. Of the 33 “native” GMDs, one identified with Thauera sp. MZ1T strain MZ1 (99.9% sequence identity at 1,409 bp). Due to the physiological relevance of Thauera sp., we sequenced our recovered Thauera (strain name SWB20) using Illumina MiSeq and PacBio, from which Illumina genome assemblies provided 300× coverage, while PacBio genome assemblies provided 159× coverage. The hybrid assemblies yielded 21 contigs for a near-complete 4,927,396 bp genome with 66.2% G+C content. Only contigs greater than 14,000 bp from the final assembly were used for subsequent annotations, which showed 4,421 protein-encoding genes and 11 rRNAs. Genomic phylogeny of SWB20 against six previously published Thauera genomes found that SWB20 clusters with Thauera aminoaromatica strains MZ1T and S2. Our SNP analysis found SWB20 having 0.66% and 0.58% SNP composition with strains MZ1T and S2, respectively, while S2 had 0.41% SNP composition when compared to MZ1T, thereby supporting the close phylogenetic topography of these three Thauera strains. Both the close phylogeny and very low SNP differences to MZ1T and S2 suggest that SWB20 is a unique T. aminoaromatica strain. Compared to the remaining four Thauera sp. genomes, SWB20 averaged ~10% SNP composition.

GMDs offer a relatively rapid method to obtain sufficient genomic template for improved genome assemblies. When using “native” growth medium to co-cultivate a complex community, GMDs with SCG recovers more bacterial diversity than from defined medium or traditional cultivation techniques, thereby providing greater access to a wider breadth of genomes of rare and relevant bacterial representatives.

Long Range Applications from Short Read Sequencing

Wednesday, 27th May 15.30 - La Fonda Ballroom - Tech Talk

Rob Tarbox[1]

[1]**10X Genomics**

The GemCode Platform is a molecular barcoding and analysis suite that delivers structural variants, and haplotypes, and other valuable information via targeted, exome, and whole genome sequencing. We will describe the platform in more detail and showcase several example uses of the platform.

Automatic Closing and Finishing of Genomes with Long Mate Pair NGS Libraries

Wednesday, 27th May 15.45 - La Fonda Ballroom - Tech Talk

David Mead[1]
[1]**Lucigen Corp.**

Long repetitive DNA sequences are abundant in most species, which creates technical challenges for the de novo assembly of even small genomes using short read next generation sequencing (NGS) methods. The incorporation of long span mate pair reads could dramatically improve the success of de novo assembly and closing of genomes by linking contigs. Existing methods are limited to 5-6 kb mate pairs, which is inadequate for most microbial or complex genomes. A new NGS library method that generates user defined mate pairs (MP) up to 100 kb has been developed. A unique barcoding strategy is used to distinguish true mate pairs from false chimeric junctions, reducing the fraction of misassembled contigs. We report the closing and finishing of four bacterial genomes using a single 10-20 kb mate pair library in conjunction with a conventional 600 bp paired end fragment library using Illumina sequencing chemistry. Genomes representing diverse sizes and %GC content were closed and finished with this simple strategy including Thermus aquaticus (2.2 Mb, 68% GC), Staphylococcus aureus (2.8 Mb, 32% GC), a Streptomyces spp. (8.6 Mb, 71% GC), and a Nonomurea spp. (10.3 Mb, 70.4% GC). SPAdes genome assembler software was able to "automatically" close four microbial genomes and finish two of them with manual review. Recent results indicate that the technology is scalable to 100 kb MP libraries, with important consequences for assembling repeat rich, complex genomes from fungi, mitochondria, chloroplasts, plants and animals. We also report on the scaffolding of human, maize, switchgrass, and a sorghum mitochondrial genome with 20-100 kb mate pair libraries. The ability to construct and sequence mate pair libraries up to 100 kb (BAC-sized paired end reads) without physical cloning simplifies the accurate closing and finishing of complex genomes economically.

Reducing Bias in Small RNA-Sequencing

Wednesday, 27th May 16.00 - La Fonda Ballroom - Tech Talk

***Masoud Toloue*[1]**

[1]*Bioo Scientific*

High throughput short read sequencing technology is ideal for the study of small RNAs, as it allows precise measurement of closely related small RNAs and novel small RNAs that hybridization-based methods like microarray and qPCR cannot achieve. Unfortunately, NGS approaches for small RNA analysis are not without their own challenges. Several studies have now shown entire datasets, including those in miRBase, to contain severe sequence bias; specifically, small RNA expression that is not accurately represented by sRNA-seq. Significant effort has gone into identifying the cause of this misrepresentation, and it is now generally accepted that bias in sRNA-seq libraries is primarily introduced during the adapter ligation steps in library preparation. Specifically, RNA ligases show sequence-specific preferences toward certain adapter-small RNA pairs, resulting in preferential inclusion of some small RNAs in sRNA-seq libraries, at the expense of others. Simply using two different adapter sequences during ligation can result in up to 30-fold differential expression for some microRNAs.

Studies have shown that ligation bias can be greatly reduced by using adapters with 2-4 random nucleotides at the ligation junctions. Thus, our approach to overcoming ligation bias in sRNA-seq libraries involves using a pool of adapters with 4 random bases at the ligation sites. Using our randomized adapter strategy, small RNA libraries were prepared and sequenced using both synthetic small RNAs and total RNA isolated from various human tissues as starting material. The results clearly demonstrate the vastly reduced bias achieved through the use of adapters with randomized ends, and show that this kit is effective in preparing small RNA libraries using total RNA from various tissues. These results demonstrate that our new streamlined small RNA-seq protocol is ideal for those needing to accurately assess small RNA abundance in diverse sample types.

A novel in vitro method for obtaining high-quality, long-range genomic information

Wednesday, 27th May 16.15 - La Fonda Ballroom - Tech Talk

Nicholas Putnam[1], Jonathan Stites[1], Brendan O'Connell[1], Brandon Rice[1], Charles Sugnet[1], Andrew Fields[1], Paul Hartley[1], David Haussler[2], Daniel Rokhsar[3], Richard Green[2]

[1]Dovetail Genomics, LLC, [2]University of California Santa Cruz, [3]University of California Berkeley

Since the beginning, the field of high-throughput sequencing has struggled to devise a fast, cheap, and reliable way to produce high-quality, long-range genomic information. We present an inexpensive, completely in vitro method for constructing sequencing libraries whose inserts span all distances up to the size of the input DNA. This library method, which involves in vitro chromatin assembly to condense DNA, requires only a few micrograms of naked DNA, no expensive equipment or exotic reagents, and can be done in a few days on a standard sequencing platform with no additional equipment.

A novel, streamlined NGS library preparation workflow employing enzymatic fragmentation results in significant improvements to library yields and sequence quality

Wednesday, 27th May 16.30 - La Fonda Ballroom - Tech Talk

Jennifer Pavlica**[1], Maryke Appel[1], Bronwen Miller[1], Victoria van Kets[1], Beverley van Rooyen[1], Heather Whitehorn[1], Martin Ranik[1], Piet Jones[1], Adriana Geldart[1], Eric van Der Walt[1]

[1]Kapa Biosystems

Continuous improvements to library preparation for next-generation sequencing (NGS) are necessary to achieve the highest data quality, and to simplify data analysis. One of the critical steps in many library preparation workflows is fragmentation of input DNA, which is accomplished through either mechanical or enzymatic means. Mechanical DNA fragmentation methods are difficult to scale or automate, and require large investments in expensive instrumentation. Current enzymatic solutions for DNA fragmentation are highly sensitive to input amount, provide poor control over fragment length distribution and exhibit sequence bias resulting in suboptimal data quality.

To address these challenges, we have developed the KAPA HyperPlus Library Preparation Kit, which provides a highly efficient, one-tube enzymatic fragmentation and library construction workflow that is compatible with a wide range of samples, inputs and sequencing applications, and is easy to automate.

In this study, we compared the KAPA HyperPlus workflow to alternative fragmentation and library construction solutions for microbial whole genome sequencing (WGS). Using microbial model organisms with extreme genomic GC content, we show that the novel enzymatic fragmentation reagent allows for robust and reproducible fragmentation across a range of inputs, and exhibits minimal sequence bias. Compared to workflows employing Covaris shearing, the KAPA HyperPlus workflow results in higher library yields, thereby reducing the requirement for library amplification. Low-bias amplification with our evolved KAPA HiFi DNA Polymerase results in higher and more uniform sequence coverage. The KAPA HyperPlus workfow exhibits several benefits over tagmentation-based workflows. These include flexibility with respect to DNA input, adapter and barcode designs and PCR-free workflows, better fragmentation control and consistency, and significant improvements to sequence coverage and coverage uniformity. As such, the KAPA HyperPlus workflow combines the speed, convenience and throughput of tagmentation-based workflows with the flexibility, control and superior sequence data quality achievable with in workflows employing mechanical fragmentation.

Oligonucleotides for Hybridization-Based Target Enrichment

Wednesday, 27th May 16.45 - La Fonda Ballroom - Tech Talk

John Havens[1]
[1]***IDT DNA***

AmpliSeq: targeted, high sensitivity sequencing for tumor mutation detection at 0.5%

Wednesday, 27th May 17.00 - La Fonda Ballroom - Tech Talk

Yongming Sun[1], Dumitru Brinza[1], Dalia Dhingra[1], Rajesh Gottimukkala[1], Charles Scafe[1], Richard Chien[1], Vidya Venkatesh[1], Kelli Bramlett[1], Fiona Hyland[1]
[1]Thermo Fisher Scientific

NGS-based oncology genomic assays for cancer mutation detection have great promises to improve cancer therapy. Ion AmpliSeq™ technology delivers simple and fast library construction for affordable targeted sequencing of specific genes or genomic regions. It requires as little as 10 ng of input DNA and compatible with FFPE samples. We have developed applications using AmpliSeq in a variety of ways, including low frequency (0.5%) cancer mutation detection, gene fusion detection (Oncomine® Gene Fusion panel), AmpliSeq pharmacogenomics panel, AmpliSeq for mutation detection in cell-free DNA (cfDNA) and circulating tumor cell (CTC).

Here we describe cancer mutation detection with various AmpliSeq panels. We focus on an analysis algorithm, using statistical modeling of next generation sequencing reads, and optimizing parameters and filters to enable sensitive and specific detection of somatic mutations to 0.5% allele ratio.

We multiplexed samples with germline blood cells, cfDNA, and CTC from same individual and sequenced these samples on one single Ion PGM 318 chip using Ion AmpliSeq CHPv2 cancer hot-spot panel. This panel allows interrogation of ~2000 relevant biomarkers from COSMIC and drug-actionable databases, and de-novo variant detection at ~20,000 genomic positions. Additionally we tested limits of variant detection with a dilution series samples with AcroMetrix® Oncology Hotspot Control DNA. The Acrometrix sample contains ~500 common cancer mutations from COSMIC and drug actionable databases. We achieved >99% sensitivity and specificity for variants present at frequency above 0.5%. Next, we spiked CTC cells from cancer cell lines into normal blood samples at ratio 1:1,000,000, obtaining 40% purity of CTC after enrichment, and demonstrated > 99% sensitivity and specificity of variant detection. We then performed analytical validation of variant detection performance with cfDNA using a dilution series of two normal blood samples, and we were able to detect in all 20 variants present in either sample with frequency above 0.5%.

To test gene fusion detection, we constructed a cell line control consisting of 5 fusion isoforms including ALK, RET and ROS fusions. We diluted this into a wild type cell line. Repeated testing of dilution ratios between 1:100 and 1:1000 on the Ion Torrent PGMTM resulted in sensitivity of 97% and specificity of 100%. At dilution ratios between 1:1 and 1:100, the sensitivity was 100%.

A technology platform for microbial genomics

Wednesday, 27th May 17.15 - La Fonda Ballroom - Tech Talk

Sam Minot[1]
[1]**One Codex**

Next-generation sequencing (NGS) is rapidly undergoing a transition from a research to an applied technology. With this shift, there is a pronounced need for analytical platforms that can scale to handle massive volumes of data and yet remain approachable for the applied non-expert end-user. This presentation will discuss a number of related areas: 1) a k-mer-centric approach to both metagenomic classification and strain-typing (which recently won the CDC's "No-Petri-Dish" Challenge); 2) the requirements and challenges in building usable, scalable software systems for metagenomic biothreat detection and biosurveillance; and 3) how One Codex is leveraging industry best practices and methodologies to enable the routine application of NGS to microbial metagenomics with scalable, repeatable, and robust bioinformatics.

Structural Variant Detection with Single Molecule Solid-State Detectors

Wednesday, 27th May 17.30 - La Fonda Ballroom - Tech Talk

John Oliver[1], Jennifer Davis[1], Tony Forget[1], Michael Kaiser[1], Jay Sage[1], Leah Seward[1], Tony Shuber[1]
[1]Nabsys Inc.

Structural variants are difficult to detect with short read technologies because genomes are hard to assemble: polymorphism, repeats, and sequencing bias can turn even small genomes into assembly nightmares. Maps constructed from long reads, however, can be used very effectively to inform the assembly process.

Solid-state, electronic nanodetectors can generate long range mapping information from single molecules that are hundreds of kilobases in length. DNA is translocated through nanochannels and detected electronically. These molecules are tagged at specific locations and those locations mapped at higher resolution than is possible with optical methods. Reads are assembled with high efficiency and used to generate accurate reference maps for genomes. Long-range information is preserved so structural rearrangements and duplications are easily identified.

Examples of small genome assemblies and the detection of structural variants in a breast cancer cell will be shown. The technology is highly scalable with the potential for much higher throughput by placing multiple detectors on each semiconductor-based chip, making rapid analysis of large, complex genomes possible.

Poster Presentations & Meet and Greet Party

630pm – 930pm, May 27th

Sponsored by Roche Diagnostics

Enjoy!!!

Characterization of the fecal microbiota in infants with botulism

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.01

Brian Shirey[1], Jan Dykes[1], Carolina Luquez[1], Susan Maslanka[1], Brian Raphael[1]
[1]National Botulism Laboratory Team, Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention

The National Botulism Surveillance System reported a yearly average of 97 cases of infant botulism in the United States (2001 – 2012) representing 68.5% of all botulism cases. Infant botulism is caused by the colonization of botulinum neurotoxin producing clostridia (BTPC) with subsequent in vivo production of botulinum neurotoxin (BoNT) in the gut of individuals less than one year old. Infant botulism has occurred with Clostridium botulinum, Clostridium baratii, and Clostridium butyricum colonizations; however, the illness is not contingent on the presence of these organisms alone.

We employed 16S rRNA gene profiling to characterize the fecal microbiota in 14 coded stool samples submitted to the CDC from patients suspected of infant botulism. Of the samples tested, 8 were from laboratory confirmed infant botulism cases, and 6 were from non-confirmed cases. Of the confirmed cases, 2 were caused by BoNT serotype A, 5 by serotype B, and 1 by serotype F. Infant ages at the time of sample collection ranged from 14 days to 350 days. Genomic DNA (gDNA) extracted from each stool sample was prepared for sequencing using the Ion Torrent platform.

Seven bacterial phyla were identified among all samples. Proteobacteria and Enterbacteriaceae abundances were significantly higher in confirmed samples, while Firmicutes abundance and the abundance ratio of Firmicutes/Proteobacteria remained significantly lower. C. botulinum and C. baratii were identified in low relative abundances in both confirmed and non-confirmed samples based on 16S rRNA gene profiles; however, BoNT gene presence cannot be verified using this approach.

Whether these differences preceded botulism or were the result of illness is unclear; however, this study provides a solid foundation for follow-up investigations. While this study demonstrates 16S rRNA gene profiling is not suitable for infant botulism detection, characterization of the fecal microflora in infants with botulism may help develop a more in depth understanding of this disease.

The findings and conclusions in this presentation are those of the author and do not represent the official position of the Centers for Disease Control and Prevention.

Improving Epidemiological Study of HIV by High Throughput Whole Genome Sequencing

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.02

***Cassie Redden**[1], **Adam Armstrong**[2], **Kimberly Bishop-Lilly**[1], **David Brett-Major**[3], **Regina Cer**[1], **Kenneth Frey**[1], **Gustavo Kijak**[4], **Vishwesh Mokashi**[5], **Gregory Rice**[1], **Paul Scott**[6], **J. Enrique Herrera-Galeano**[1]*

[1]Henry M Jackson Foundation/Naval Medical Research Center, [2]Navy Bloodborne Infection Management Center, [3]Uniformed Services University of the Health Sciences, [4]Henry M Jackson Foundation/Military HIV Research Program,

Traditionally, epidemiological studies of HIV-1 have focused on two regions of the viral genome: pol and gag/env. Portions of these regions are often amplified and sequenced as part of clinical management of individual patients, and yet the regions sequenced only constitute roughly a tenth of the viral genome. However, a protocol to capture nearly full length viral genomes using archived plasma samples and next generation sequencing (NGS) could greatly improve the epidemiological study of this virus. To investigate whether high throughput whole genome sequencing could improve HIV epidemiology, RNA was extracted from HIV positive frozen plasma samples and reverse transcribed. Multiple overlapping regions were then PCR amplified using a previously published protocol. Of the eight samples tested, seven had amplification in at least one region and four had amplification in all regions and this latter subset was chosen for sequencing. These amplicons were then combined, indexed Nextera XT libraries were created and multiplexed sequencing completed on the Illumina MiSeq. Following this procedure, over 8,000 base pairs (bp) of the roughly 10,000 bp genome was produced for the four samples sequenced. Based on these results, HIV-1 positive plasma samples that have been archived and stored could be used to generate NGS data for nearly full length genomes. The resulting data, studied epidemiologically, could have a major impact on the current understanding of the transmission of this disease.

Amplicon based 16S ribosomal RNA Sequencing and Genus Identification

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.03

***Dawn Obermoeller*[1]*, Masoud Toloue*[1]*, Jan Risinger*[1]**

[1]*Bioo Scientific*

Next generation sequencing analysis of 16S ribosomal RNA (rRNA) is commonly used to identify bacterial species and perform taxonomic studies. Bacterial 16S rRNA genes contain 9 hyper-variable regions with considerable sequence diversity among different bacterial species and can be used for species id. Rapid determination of highly complex bacterial populations through targeted amplification can provide an accurate gauge of diversity at taxonomic hierarchies as low as the genus level. A single 16S rRNA hyper-variable domain does not have enough sequence diversity to distinguish genera. With increased read lengths of MiSeq chemistry, Bioo Scientific has expanded the common analysis of the fourth hyper-variable domain (V4) of prokaryotic 16S rRNA to V1, V2 and V3 regions simultaneously. Optimized preparation through a streamlined standardized procedure allows for high-quality, reproducible libraries. This optimization can be applied to different windows of 16S rRNA as well as other relevant prokaryotic taxonomic markers.

Gene Conversion and Cotton Allotetraploid Evolution

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.04

Justin Page[1], Joshua Udall[1]
[1]Brigham Young University

Cotton tetraploids are a valuable source of fiber for textiles and their genome evolution has not been fully characterized. We analyze the evolution of allotetraploid cotton using whole-genome resequencing data consisting of over 18 billion reads from over 30 lines and all 7 allotetraploid species. We mapped the polyploid reads to both the G. raimondii and G. arboreum reference sequences using GSNAP and PolyCat, eliminating or severely reducing mapping biases between species with information from ~25 million homoeo-SNPs. With this depth and breadth of data, we examine non-reciprocal homoeologous recombination (gene conversion). Past studies have been unable to confidently distinguish between gene conversion events and polymorphisms arising in diploid lines. With the broad sampling of our study, we show that roughly half of putative events reflect true non-reciprocal homoeologous recombination events. We provide a refined phylogeny based on whole-genome SNP data, including the AT and DT genomes of each species--including the newly described G. eckmanianum and putative AD7 species--and 10 extant diploid relatives. We identify introgression of G. barbadense into G. hirsutum and vice versa. Homoeo-SNPs between the A and D genomes and allele SNPs within and between tetraploid species are made available for free use at CottonGen.

Hybrid assembly of Azuki bean using Illumina and PacBio sequence data as input for the CLC scaffolding tool

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.05

Mari Miyamoto[1], Hiroaki Sakai[2], Ken Naito[2], Leif Schausser[3], Martin Simonsen[3], Arne Materna[3]

[1]QIAGEN KK, [2]National Institute of Agrobiological Sciences, [3]QIAGEN Aarhus A/S

Genome assembly using next generation sequence (NGS) data has become a common technique. However, obtaining high quality assemblies still remains difficulties due to genome complexity, as caused by genomic repeat regions that cannot be easily resolved via the assembly of short reads only. The emergence of SMRT sequencing technologies, as implemented by Pacific Biosciences (PacBio), delivering long reads that originate from single DNA molecules, has greatly improved the ability to resolve repetitive regions. Specifically Pacific PacBio data have become the standard for assembling smaller genomes. And while evidently PacBio data can also improve the assembly of higher order organisms it is not always cost efficient to rely on this data type alone.

We present the CLC scaffolder (“Join Contigs” tool, part of the CLC Genome Finishing Module), which enables hybrid assembly using Illumina contigs and raw (not error corrected) PacBio reads. The Join Contigs tool is made available through the CLC Genome Finishing Module, an add-on to the CLC Genomics Workbench (now part of the QIAGEN Bioinformatics product portfolio). The Join Contigs tool examines if the PacBio reads are mapped to the Illumina contigs then estimates if two contigs should be joined. Furthermore quality is assessed and low quality regions are identified.

To validate our tool, we assembled Vigna angularis (Azuki bean), which was sequenced on PacBio and Illumina sequencers. Azuki bean is widely cultivated in Northeast Asia. The genome size of the diploid crop Vigna angularis is 538Mb. Illumina reads were assembled using the CLC de novo assembler and the CLC scaffolder to join the Illumina contigs and PacBio reads. Adding PacBio data greatly improves the length of contigs. We also show the result of a variety of conditions that affected the quality of assembly and scaffolding.

Using PacBio to Improve Metagenomes

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.06

***Alicia Clum*[1], *Alex Copeland*[1]**

[1]*Joint Genome Institute*

The decrease in sequencing cost in recent years has resulted in expansions in the field of metagenomics. Often assemblies from these datasets are highly fragmented and incomplete. We will discuss how PacBio data can be applied to improve these assemblies and how base modification data can be used to validate genome bins.

RNAseq-Based Analysis of Clinical Influenza Samples from the Republic of Georgia

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.07

***Adam Kotorashvili*[1], Nato Kotaria[1], Anna MacHablishvili[1], Jason Farlow[2], Cheryl Gleasner[3], Momchilo Vuyisich[3], Tracy Erkkila[3]**

[1]NCDC/Lugar Center, [2]Farlow Scientific Consulting Company, LLC, [3]Los Alamos National Laboratory*

Influenza A viruses are members of the Orthomyxoviridea family, which comprises enveloped, negative-sense, single-stranded RNA, viruses containing a genome divided over eight RNA segments. Influenza virus type A is subdivided based on the antigenic properties of the major surface glycoproteins: hemagglutinin (HA) and neuraminidase (NA). To date, 17 HA and 10 NA subtypes have been found in nature.

Influenza clinical samples were collected across the country of Georgia under Influenza surveillance program by National Centre for Disease Control and Public Health. To identify genes from Influenza positive clinical samples, RNA was extracted from nasopharyngeal swabs in transportation medium (following NAMRU-3 (U.S. Naval Medical Research Unit #3) and WHO guidelines for Influenza virus). We performed whole genome sequencing of pathogen RNA recovered from four PCR positive samples from different regions of Georgia.

Illumina MiSeq sequencing resulted in the generation of 4-5 million 600 bp reads per sample. Sequence read mapping analysis was performed using CLC Bio software. Multiple DNA alignment and phylogenetic analyses of antigenic hemagglutinin (HA) and neuroaminidase (NA) genes were conducted. Preliminary analyses of major surface glycoproteins (HA and NA) suggested that the virus lineage circulating in Georgia exhibits the closest sequence similarity to Influenza A virus reported in Singapore: Influenza A virus (A/Singapore/H2013.529/2013 (H3N2)) segment four hemagglutinin (HA) gene and (A/Singapore/H2013.30a/2013 (H3N2)) segment four hemagglutinin gene. The HA gene shared the highest identity with Singapore/H2013 (H3N2), while the NA gene was most closely related to Influenza A virus reported in Pennsylvania (A/Pennsylvania/12/2013(H3N2)) segment 6 neuraminidase (NA) gene.

Genome Sequencing for Two Single-Chromosome Vibrio cholerae Isolates, Strains 1154-74 (Serogroup O49) and 10432-62 (Serogroup O27)

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.08

***Gary Xie**[1], **Karen Davenport**[1], **Shannon Johnson**[1], **Patrick Chain**[1], **Shanmuga Sozhamannan**[2]*

[1]Los Alamos National Laboratory, [2]CRITICAL REAGENTS PROGRAM

The complete genome sequences of Vibrio cholerae isolates, strains 1154-74 (Serogroup O49) and 10432-62 (Serogroup O27) were determined. The initial sequence analysis of these non-O1/non-O139 isolates revealed pervasive genetic and genomic structural diversity, including indels, duplications, and fusions of the usual two chromosomes in these two genomes. This is the first report on naturally occurring Vibrio cholerae strains that have this unusual genomic topology of a single chromosome as opposed to the normal paradigm of two chromosomes. Within both genomes, prophage elements were identified flanking the insertion sites that may explain the mechanism of chromosome fusions.

Optical genome mapping as a tool to aid genome assembly: Process overview and results from applying it to the assembly of mouse genomes

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.09

***Matthew Dunn**[1], Michelle Smith[1], Jonathan Wood[1], Michelle Dignam[1]*
[1]Wellcome Trust Sanger Institute

Optical based genome mapping is a technique now commercialized with different approaches by a number of biotech companies. It is a process whereby long single DNA molecules are labelled or alternatively digested with restriction enzymes, linearized, viewed, analysed and subsequently assembled into a high definition map, providing a long range structural view of the genome and it's architecture.

Utilising the Argus system from OpGen and the Irys system from BioNano Genomics we have taken two approaches to mapping mouse genomes. Firstly, we have utilised OpGen's novel Genome Builder process. Genome Builder uses local single molecule assemblies from optical mapping to join sequence contigs together, creating large sequence scaffolds. The resultant local optical map assemblies span several megabases in size, providing previously absent long-range information. We used this information to join sequence contigs, size gaps and identify regions of misassembly.

Utilising the Irys system from BioNano Genomics we have been able to generate entirely de novo mouse assemblies, by analysing this data in IrysView software the mapping information was utilised to orientate sequence contigs and size gaps by bridging across repeats and other complex elements that can break NGS assemblies. The data also allowed identification and resolution of regions of sequence misassembly.

These approaches have yielded rapid and significant improvements in the genome assemblies which further aids the downstream analysis and understanding of these genomes. The methodology and results of both these approaches are outlined across several mapping projects.

Forensic Science Research and Development Funding Program at the National Institute of Justice

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.10

***Minh Nguyen*[1]**

[1]*National Institute of Justice*

The National Institute of Justice (NIJ) - the research, development and evaluation agency of the U.S. Department of Justice - is dedicated to improving knowledge and understanding of crime and justice issues through science. NIJ's Office of Investigative and Forensic Sciences supports this mission by sponsoring research to provide objective, independent, evidence-based knowledge and tools to meet the challenges of criminal justice, particularly at the State and local levels.

As next generation sequencing (NGS) technologies have advanced, the interest in developing NGS based methods and evaluating NGS workflows for application to forensic issues has increased. In fiscal year 2014 alone, NIJ invested over four million dollars in research and development projects to evaluate NGS technologies, to develop/optimize NGS based methods, and to analyze DNA sequences generated with NGS technologies. Projects vary from DNA sequence analysis for individual identification, phenotypic information, information about the type or age of a biological stain, the analysis of microbiomes as trace evidence, and molecular autopsy to determine causes of death. NIJ anticipates continued interest in NGS technologies for forensic applications and is interested in engaging the genomics community in examining forensically relevant research questions.

This poster will present an overview of NIJ's research and development portfolio, information on its funding cycle, and general information about research and development funding opportunities at NIJ.

Highly multiplexed amplicon sequencing provides rapid, sensitive detection and genetic characterization of Burkholderia pseudomallei

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.11

***Josie Delisle*[1], Jim Schupp[1], Jason Sahl[1], Rebecca Colman[1], Hannah Heaton[1], John Gillece[1], Adam Vazquez[2], Carina Hall[2], Joseph Busch[2], Mark Mayo[3], Bart Currie[3], David Engelthaler[1], Paul Keim[1], Dave Wagner[2]**

[1]Translational Genomics Research Institute, [2]Northern Arizona University, [3]Menzies School of Health Research

Rapid detection and characterization of clinical and forensic materials suspected of containing Burkholderia pseudomallei, a public health and potential bioterrorism agent endemic to Southeast Asia and Northern Australia, would be of enormous benefit to epidemiological and forensic investigations. Current methodologies, such as real time PCR, allow rapid detection but only limited characterization. Next generation sequencing of multiple informative genetic loci can provide efficient, rapid detection and differentiation from near neighbor species, as well as fine scale genetic characterization. We have developed a 28 locus amplicon sequencing system that results in 1) detection of B. pseudomallei; 2) differentiation from B. mallei and near neighbor species; 3) potential detection of strain mixtures; 4) differentiation within B. pseudomallei; and 5) virulence gene characterization (10 vir genes), within 24-48 hours, and from both culture and complex environmental or clinical sample material. The system couples highly multiplexed amplification reactions with a universal amplicon indexing system, resulting in efficient multilocus amplicon sequencing from potentially hundreds of samples in a single Illumina MiSeq sequencing run, with sequence results in as little as 24 hours. Utilizing redundant targets identified with Blast Score Ratio analysis for species identification, we show virtually 100% specificity using a panel of B. pseudomallei, B. mallei and close near neighbors, such as B. thailandensis, B. oklahomensis, and B. humptydooensis, among others. We also demonstrate differentiation within B. pseudomallei strains, utilizing variation within the targeted species specific, MLST and virulence loci.

Assessing congruence in SNP genotypes determined from CLC Genomics Workbench and Life Technologies HID SNP Genotyper

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.12

James Robertson[1], Kelly Meiklejohn[1]
[1]FBI Academy / FBI Laboratory

De novo sequence assembly of data obtained from massively parallel sequencing (MPS) platforms can be computationally intense, and differences in the underlying algorithm can impact the resulting alignment. Generally with the purchase of a MPS, the manufacturer provides software specifically designed for their platform that facilitates sequence alignment and data analysis. There are however a number of stand alone software programs available either commercially or via open source (e.g. CLC Genomics Workbench [Qiagen] and NextGENe® [SoftGenetics], GeneTalk Basic [GeneTalk]), that have been developed specifically for MPS data analysis from a range of platforIn this study we assessed whether the genotypes of the 124 SNPs included in Life Technologies HID-Ion Ampliseq Identity Panel (ThermoFisher Scientific) were sensitive to variation in alignment algorithUsing data collected for the panel from a range of purchased pure native DNAs and forensic type samples, we compared the calls from the manfacturer's HID SNP Genotyper to those obtained from the CLC Genomics Workbench. In addition to this, the 'quality based variant detection' tool was run in CLC to establish if all the SNPs included in the Identity Panel were independently identified, and whether any additional potentially informative SNPs were present in the sequenced regions.

Graph Based Data Structures & Algorithms for Pan-Genomics

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.13

Thiruvarangan Ramaraj[1], Joann Mudge[1], Brendan Mumey[2]
[1]National Center for Genome Resources (NCGR), [2]Montana State University-Bozeman

Single Nucleotide Polymorphisms (SNPs), Insertions and Deletions (INDELs), and Structural Variations (SVs) are the basis of genetic variation among individuals and populations. Second and third generation high throughput-sequencing technologies have fundamentally changed our biological interpretation of genomes and notably have transformed analysis and characterization of the genome-wide variations present in a population or a particular environment. As a result of this revolution in the next generation sequencing technologies we now have a large volume of genome sequences of species that represent major phylogenetic clades. Having multiple, independent genomic assemblies from a species presents the opportunity to move away from a single reference per species, incorporating information from species across the phylogenomic range of the species into a pan-genomic reference that can better organize and query the underlying genetic variation. Tools have started to explore multiple genomes in bioinformatics analyses. Several tools have evolved to take advantage of information from multiple, closely related genomes (species, strains/lines) to perform bioinformatics analyses such as variant detection without the bias introduced from using a single reference. In this project we consider challenges and opportunities that arise from pan-genomics. Specifically we will investigate improvements to graphical data structures used to represent pan-genomes and associated algorithms for analyzing, representing, and visualizing pan-genomes to ultimately investigate the true level of genomic variation through deep sequencing, de novo assembly and comparative analysis.

NGS Data Analysis of Bacillus anthracis and Francisella tularensis - Comparison of Two Bioinformatic Tools

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.14

Ekaterine Khmaladze*[1], *Gvantsa Chanturia*[1], Ekaterine Zhgenti[1], Tracy Erkkila[2], Patrick Chain[2], Karen Davenport[2], Chien-Chi Lo[2], Po-E Li[2], Sanaa Ahmed[2]

[1]National Center for Disease Control and Public Health/Lugar Centre, Tbilisi,, [2]Los Alamos National Laboratory

Bacillus anthracis cause the acute fatal disease anthrax and is a proven biological weapon. It is endemic in Georgia and displays subspecies-specific differences in genetic diversity. The B. anthracis genome consists of an approximately 5.3-Mb chromosome and two plasmids, pXO1 (182 kb) and pXO2 (96 kb). Francisella tularensis subspecies holarctica is also endemic pathogen and poses biological threat for the country. In contrast to B. anthracis, it has a small (1.8-Mb) genome without plasmids but rich with IS elements.

In order to study Georgian genetic variants of these organisms, 10 B. anthracis and six F. tularensis strains from the live culture repository of NCDC were chosen from different Single Nucleotide Polymorphism (SNP) subclades and Multiple-Locus Variable number tandem repeat Analysis (MLVA) genotypes that were recently discovered based on genetic characterization of Georgian isolates.

For this purpose, DNA fragment libraries were generated from genomic DNA according to Illumina next-generation sequencing sample preparation method. B. anthracis DNAs were shredded by nebulisation, while Covaris instrument has been applied for fragmentation of F. tularensis genomic DNAs. The final size with the average of ~450bp of the prepared libraries was determined by Agilent 2100 Bioanalyzer. Sequencing was performed using Illumina 300 cycle sequencing kit on the MiSeq platform at NCDC Lugar Center in Tbilisi, Georgia. Obtained raw data of 150bp length reads were analyzed using two different software suites available at the Lugar Center - CLC Genomics Workbench and EDGE Bioinformatics. Assemblies were aligned to the closest to SNP subclade reference genomes - Ames ancestor and Sterne for B. anthracis and Live Vaccine Strain (LVS) for F. tularensis.

Similar trimming and assembly parameters were set for analyzing workflows with both software: minimum read length – 50bp, minimum contig size – 200bp, for CLC the word size (Kmer size) was set to its maximum - 64 bp, while in EDGE minimum Kmer size was defined as – 31, maximum – 128, step size - 20bp.

The trimming results and average fold coverage on appropriate reference genomes for both organisms were similar, as well as the size and number of contigs after performing a de novo assembly.

The overall outcome from both packages - commercially available CLC Genomics and open source EDGE Bioinformatics revealed comparable results, providing confidence that the open source tools are equally valid; CLC is windows based, powerful and flexible software, while EDGE is easy to operate multifunctional package of different software with many useful tools.

Characterization Of Passive Immune Transfer Thoughout Lactation In A Model Marsupial

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.15

Bethaney Fehrenkamp[1], Victoria Hansen[1], Robert Miller[1]
[1]University of New Mexico

Lactation is a complex strategy utilized to provide continued nutritional and immunological support to developing offspring outside of the womb. This is especially true for marsupials, a lineage of mammals with a short gestation, and limited placental development. Marsupial neonates are born highly altricial and have an increased reliance on an extended lactation program. This early exposure to pathogens, prior to the development of a functional immune system, requires a complex immune investment throughout lactation. Previous research in Australian marsupials has shown two distinct waves of immune transfer correlating with offspring suckling behavior and potential exposure to pathogens. High-throughput RNA sequencing and bioinformatics analysis was utilized to characterize immune related gene expression in the mammary tissues of a model marsupial, Monodelphis domestica, throughout the course of lactation. Results are consistent with migration of lymphocytes into the mammary tissue at two crucial periods of neonatal development, immediately following birth and again as the offspring is preparing to wean, and correlate with potential exposure to pathogens. Expression of immunoglobulins also follows this pattern of expression, excepting for IgA, which displays a more consistent pattern of elevated expression throughout lactation. This expression pattern may be related to establishment and regulation of the developing gut flora of the offspring. Results provide further evidence for the ancient role of lactation providing immune protection to developing offspring.

The Exome Coverage and Identification (ExCID) Report: a gene survey tool for WES and WGS applications

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.16

Christian Buhay[1], Rashesh Sanghvi[1], Qiaoyan Wang[1], Kimberly Walker[1], Harsha Doddapaneni[1], Jianhong Hu[1], Mark Wang[1], Yi Han[1], Huyen Dinh[1], Eric Boerwinkle[1], Donna Muzny[1], Richard Gibbs[1]

[1]Baylor College of Medicine

The Exome Coverage and Identification (ExCID) Report was developed at the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) to represent gene transcript and exon sequence coverage for samples analyzed with the VCRome Whole Exome sequencing (WES) reagent. Since March 2013, the report has been used to analyze more than 18,000 WES research samples for the BCM-HGSC and more than 3000 WES clinical samples at our Whole Genome Laboratory (WGL).

ExCID assesses target sequence coverage, annotates targets with gene, transcript and exon information and reports intervals below 20X. In addition, ExCID has batch analysis features that can compare data from hundreds of samples to reveal trends in the performance of large scale sequencing projects. Results can be visualized as ‘coverage tracks’ in popular browsers such as the Integrative Genome Browser and the UCSC Genome Browser.

ExCID has been used to survey and compare WES performance in HGSC R&D efforts. Technical developments such as capture reagent spike-in (i.e. panel killer) permit improved coverage of nearly all clinically targeted genes (N=3,200 at 100% coverage); rapid capture methods (i.e. lightning capture) have decreased the capture and hybridization time from 4 days to 8 hrs without loss of coverage or base quality, facilitating sequencing applications in the prenatal and neonatal intensive care environments. Furthermore, ExCID is also being used as the standard metric gathering tool in the Clinical Sequencing Exploratory Research Sequencing Standards working group, a consortium of nine clinical sequencing sites across the country.

ExCID is available under public license in the GitHub repository:
https://github.com/cbuhay/ExCID

PacBio based assemblies of small eukaryotes

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) – Poster 1a.17

***Yuliya Kunde*[1], *Karen Davenport*[1], *Shawn Starkenburg*[1], *Cheryl Gleasner*[1], *Olga Chertkov*[1], *Kim McMurry*[1]**

[1] *Los Alamos National Laboratory*

Photosynthetic microalgae are a promising source of feedstock material for biofuels. The mechanism(s) of lipid production are not fully understood, but the widely accepted hypothesis is that under stress conditions, microalgae convert excess energy from light into storage compounds like starch and lipids. Currently, high quality genome assemblies from microalgae production strains are not available (thousands of contigs). Access to nearly finished genomes for these organisms will significantly improve our understanding of key metabolic pathways, and inform rational genetic engineering approaches. For this purpose, genomic DNA from two top ranking candidates, Chlorella sp. (strains 1228 and 1230) and Scenedesmus obliquus, was converted into 20kb libraries for sequencing and assembly with PacBio and HGAP, respectively. The Pacbio based assemblies were further improved with short reads from Illumina or OpGen optical maps. Herein, we will present the comparisons of these assembly methods as well as cost-benefit analysis of generating hybrid assemblies with Pacbio and OpGen.

Comparative genomics between Tephritid species

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) - Poster 1a.18

***Bernarda Calla*[1], Sheina Sim[2], Scott Geib[1]**
[1]USDA-ARS Daniel K. Inouye US PBARC, [2]University of Hawaii, Manoa*

True fruit flies (Diptera: Tephritidae) constitute the most relevant family of flies in terms of agricultural importance. Many species in this family are phytophagous pests and attack fruits of commercial crops. Economically important taxa include Bactrocera spp., Anastrepa spp., Ceratitis capitata and Rhagoletis spp., and several area-wide programs are aimed to the control and eradication of these pests, costing agencies several million dollars a year. Availability of quality genomic information for these species is scarce and inconsistent. We have undertaken the task of obtaining high quality, uniformly annotated genome assemblies for these species to generate a database that allows for cross-comparison between them, and as a basis and common point for other Tephritid research. For this purpose, sequencing data is obtained from individual flies and assembled using ALLPATHS and/or DISCOVAR (Broad Institute). Parent flies as well as siblings derived from iso-crossings are sequenced in order to reduce heterozygosity. Consistent and synonymized annotation is obtained based off of orthology and with the use of the NCBI Eukariotic Genome Annotation Pipeline. Data will be publicly available utilizing the USDA-ARS i5k web portal. On the short term, we are targeting to release the genomes of approximately 10 species of Tephritid flies from throughout the phylogeny of this family to serve as a foundation for this pest family. This approach to create consistent genomic assemblies of related species can serve as a model for other important groups of organisms where comparative genomics is of interest.

Genomics Science and Technology for International Cooperative Research

Wednesday, 27th May 18.30 - Santa Fe Room (Sponsored by Roche) - Poster 1a.19

Helen Cui[1], Tracy Erkkila[1], Patrick Chain[1], Momchilo Vuyisich[1], Karen Davenport[1], Cheryl Gleasner[1], Shannon Johnson[1]
[1]Los Alamos National Laboratory

Genomic science and technologies are transforming life sciences globally, and becoming an important area for international collaboration. Los Alamos National Laboratory is leveraging a long history of genomics research, and strong experience and expertise in the field, to assist multiple international partner countries in advancing their genomics and bioinformatics capabilities. Our current partner countries and regions include the Republic of Georgia, Jordan, Kenya, Yemen, Gabon, Uganda, and South East Asia. These collaborations are primarily sponsored by the US Defense Department and US State Department, with additional support from the United Kingdom and Canadian governments. Collaborations with other countries and regions are also being initiated.

We have focused on providing a genomics-based scientific approach, including introductions to next generation sequencing technology and bioinformatic analysis of sequencing data for pathogen detection, characterization, and biosurveillance applications. We begin by assisting the host nations in developing the capabilities that are urgently needed to address pressing challenges in infectious disease outbreaks, implementing safe laboratory practices, and developing infectious disease detection and characterization techniques that can be maintained and further developed by the host countries. We continue to develop and provide the partner countries with sequencing protocols and bioinformatics pipelines that enable efficient sample preparation, instrument operation, and next generation sequencing data processing and analysis.

As part of the training and cooperative research development effort, Los Alamos has been conducting a Next Generation Sequencing techniques and bioinformatics training workshop. 2015 is the third year in the annual training workshop series. The topics of the training workshop include fundamental knowledge and practical skills of applying NGS and data analysis. The attendance has increased from 15 in 2013 to 38 in 2015, including scientists and laboratory practitioners from most of the partner countries.

Hyper-finishing: catching DNA mobilization in the act

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.01

Corey Hudson[1], Kelly Williams[1]
[1]Sandia National Laboratories

Mobile genetic elements are a major agent of rapid genome evolution in bacteria. Many chromosomal genomic elements begin their mobile stage by excising in free circular form. Genome finishing would ideally detect and report these subgenomic circles. The signatures of mobile DNAs – their circular junction sequences – can be detected by high throughput sequencing (HTseq) even at low levels. We detected circular forms of six genomic islands and two copies of an insertion sequence in a Klebsiella pneumoniae strain, that appeared upon mitomycin induction and in some cases spontaneously. While sequencing a Cupriavidus metallidurans strain, a spontaneously induced phage circle was detected that had confused assembly software into assembling it as a chromosomal tandem prophage duplicate. We have identified additional prophage tandem duplicates in public genomes, that may likewise instead be a mixture of single-copy prophage and the replicating free circular phage. To automate detection of mobile DNA activity among raw HTseq data, we developed Juxtaposer, a software that mines HTseq data to find small circularized genomic elements, as well as the genomic scars left by their deletion from the chromosome, even if at low levels. This is a tool that directly discovers actively mobilized elements, rather than merely confirmatory PCR tests. Juxtaposer relies on no bioinformatic prejudices about the features of the mobile elements it discovers. Applying this software to publicly available HTseq datasets, we found evidence for mobility of genomic islands and other classes of mobile genetic elements. We recommend routine application of mitomycin in bacterial and archaeal genome projects, to improve finishing of the mobilome.

Automated Evaluation of Structural Variant Detection Tools Using Simulated Reference Genomes

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.02

Zhanyang Zhu[1], Andy Wing Chun Pang[1], Warren Andrews[1], Ernest T. Lam[1], Xiang Zhou[1], Tiffany Y. Liang[1], Vladimir Dergachev[1], Thomas Anantharaman[1], Alex Hastie[1], Han Cao[1], Zeljko Dzakula[1]

[1]BioNano Genomics

With recent technological advances, whole genome sequencing and mapping have become much faster and cheaper, setting the stage for the development of computational tools to find structural variants (SVs) from sequencing and mapping data. However, the evaluation of these tools is hampered by the absence of a human sample with a complete annotation of SVs and by the need for a large number of samples with known ground truth to ensure adequate statistical power. To address the lack of known ground truth and to achieve high statistical power when measuring the performance of new SV detection tools, we developed an algorithm SVET (SV Evaluation Toolkit) to generate edited reference genomes. We first collected over 1,300,000 known SVs from 23 population studies sampling from over 2,600 individuals, and then incorporated over 150,000 of these events into hg19 to generate 35 edited reference genomes. On average, over 4,200 events were incorporated into each reference. We used these as reference genomes when calling SVs against assemblies of BioNano Genomics data collected on an ensemble of euploid human samples. The procedure enabled us to reliably evaluate the algorithmic accuracy of a large number of SV calls, thus achieving high statistical power when evaluating SV detection tools in spite of the relatively modest number of measured euploid samples.

We applied SVET to evaluate our SV detection tools that use BioNano whole genome mapping data. With our Irys® genome mapping platform and NanoChannel technology, we collected ultra-long DNA molecules (ranging from 150 kb to 2.5 Mb) and assembled them de novo into genome maps. In total, we assembled experimental data collected on 34 euploid human samples, and called SVs by aligning these assemblies against the public hg19 and 35 edited references.

In this presentation, we describe SVET results on insertion, deletion and inversion calls derived from alignments of experimentally measured and assembled BioNano data against edited references. We also discuss the possible applications of SVET in NGS SV validations. The simulation of reference genomes, rather than sample genomes, enables us to evaluate a variety of algorithms that detect all types of structural aberrations using only a limited number of measured samples. The SV accuracy assessment realistically reflects mapping and measurement errors and allows for accurate tally of false calls without the need for a fully-annotated human SV "gold standard."

An Amplicon Sequencing Analysis Pipeline for High-Throughput Characterization of Complex Samples

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.03

***Darrin Lemmer*[1], *Jolene Bowers*[1], *Erin Kelley*[1], *Elizabeth Driebe*[1], *Jim Schupp*[1], *David Engelthaler*[1], *Paul Keim*[1]**

[1]*Translational Genomics Research Institute, Pathogen Genomics Division*

A new technique in amplicon sequencing allows for multiplexing many different target amplicons for a single sample together on the same next-generation sequencing run. With the number of samples really only being limited by the number of unique indexes, it is easy to have 10,000 or more amplicons sequenced on a single MiSeq run. This technique is ideal for analyzing clinical samples, as you can run tens to hundreds of different DNA-based assays directly on each sample without culturing bacterial isolates or dealing with human DNA contamination masking the pathogen signal as you would for whole genome or metagenomic sequencing. We're using these assays for pathogen species and strain identification, and determining antibiotic resistance and other virulence factors. Interpreting the significance of the target amplicons depends on a number of factors: whether they are present or not, which gene variant is the best match, or whether there is a SNP at a specific position – or any amino acid changing SNP – within the target.

While aligning the reads to the targets is fairly straightforward, examining each alignment to determine its significance becomes challenging, particularly when you are looking at thousands of different alignments per run. Additionally, to be useful in a clinical setting, we need the results to be accurate, easy to interpret, and as fast as possible to guide appropriate antibiotic therapy. Presented here is a high-throughput analysis pipeline that includes a standardized format for defining all of your target assays, aligning your read data and interpreting the results based on your target definitions, and outputting the results in a variety of customizable formats (HTML, Excel, PDF, etc.) and levels of detail, from clinical summaries (what pathogens and antibiotic resistance genes are present in the sample) to full details including read counts and SNP positions for each target. While primarily focused on detecting antibiotic resistant pathogens in clinical samples, the standardized target definition and customizable output formats allow this pipeline to be used with any amplicon targets on any type of sample.

Towards Finished Genomes: Drosophila pseudoobscura as a case study

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.04

Shwetha Murali[1], Kim Worley[1], Doris Bachtrog[2], Stephen Schaeffer[3], Susan Brown[4], Stephen Richards[1]
[1]Baylor College of Medicine, [2]University of California Berkeley, [3]Penn State University, [4]Kansas State University

To date, there are few finished eukaryotic large genome assemblies: Human, Mouse, Drosophila, C. elegans and Arabidopsis. The eukaryotes so far have been served by low cost draft genome assemblies whose quality has declined with the advent of short reads. Tens of thousands of gaps and short contigs and scaffolds plague genome analysis in many species, especially more polymorphic species beyond the mammals. Insects produce particularly poor genome assemblies due to high polymorphism (5-50 fold greater than human), small physical size (requiring multiple individuals for DNA extraction), and usually the inability to inbreed (for genome homozygocity).

Drosophila pseudoobscura is the second Drosophila species to be sequenced – it is a 175Mb Sanger draft genome published in 2005 with a contig N50 of 52kb, and scaffold N50 of 1Mb. The species was popularized by Sturtevant and Dobzhansky, who discovered 3rd chromosome inversion polymorphisms forming stable geographic and altitudinal clines across the US. Despite a high quality draft genome, its deficiencies have hindered full characterization of all chromosome inversion polymorphisWe are now aiming to finish the D. pseudoobscura genome using as little targeted sequencing as possible while using data generated from the most current technologies and platfor

We generated four datasets with complimenting strengths and weaknesses to add to the sequencing data from the original inbred line (14 generations of sib-sib mating) and original DNA isolation: 70X PacBio long read data, 140X Illumina of different insert sizes, BioNano sequence motif mapping data, and Hi-C chromatin sequencing data. We experimented with different combinations of the datasets and assembler parameters to generate multiple assemblies and compared them. Hi-C data was used to place all contigs/scaffolds on chromosomes. We report multi-megabase contigs, assembly chimerism rates based on alignment of sequences to D. pseudoobscura and melanogaster syntenic chromosome arms, and validation using BioNano optical map information. Finally we report on our efforts to combine all these datasets into a single finished “archival quality” genome. All data is available on the HGSC website at https://www.hgsc.bcm.edu/arthropods/drosophila-pseudoobscura-genome-project and has already been used by multiple investigators to test new assembly strategies.

Automatic Closing and Finishing of Genomes with Long Mate Pair NGS Libraries

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.05

David Mead[1], Scott Monsma[1], Svetlana Jasinovica[1], Erin Ferguson[1], Brendan Keough[1], Megan Niebauer[1], Michael Lodes[1]
[1]Lucigen Corp.

Long repetitive DNA sequences are abundant in most species, which creates technical challenges for the de novo assembly of even small genomes using short read next generation sequencing (NGS) methods. The incorporation of long span mate pair reads could dramatically improve the success of de novo assembly and closing of genomes by linking contigs. Existing methods are limited to 5-6 kb mate pairs, which is inadequate for most microbial or complex genomes. A new NGS library method that generates user defined mate pairs (MP) up to 100 kb has been developed. A unique barcoding strategy is used to distinguish true mate pairs from false chimeric junctions, reducing the fraction of misassembled contigs. We report the closing and finishing of four bacterial genomes using a single 10-20 kb mate pair library in conjunction with a conventional 600 bp paired end fragment library using Illumina sequencing chemistry. Genomes representing diverse sizes and %GC content were closed and finished with this simple strategy including Thermus aquaticus (2.2 Mb, 68% GC), Staphylococcus aureus (2.8 Mb, 32% GC), a Streptomyces spp. (8.6 Mb, 71% GC), and a Nonomurea spp. (10.3 Mb, 70.4% GC). SPAdes genome assembler software was able to "automatically" close four microbial genomes and finish two of them with manual review. Recent results indicate that the technology is scalable to 100 kb MP libraries, with important consequences for assembling repeat rich, complex genomes from fungi, mitochondria, chloroplasts, plants and animals. We also report on the scaffolding of human, maize, switchgrass, and a sorghum mitochondrial genome with 20-100 kb mate pair libraries. The ability to construct and sequence mate pair libraries up to 100 kb (BAC-sized paired end reads) without physical cloning simplifies the accurate closing and finishing of complex genomes economically.

Transcriptomic and proteomic dynamics in the metabolism of a diazotrophic cyanobacterium, Cyanothece sp. PCC 7822, during a diurnal day-night cycle

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.06

***Lou Sherman*[1], *David Welkie*[1], *Xiaohui Zhang*[1], *Meng Markillie*[2], *Ronald Taylor*[2], *Galya Orr*[2], *Jon Jacobs*[2], *Ketaki Bhide*[1], *Jyothi Thimmapuram*[1], *Marina Gritsenko*[2], *Hugh Mitchell*[2], *Richard Smith*[2]**

[1]Purdue University, [2]Environmental Molecular Sciences Lab, PNNL

Background: Cyanothece sp. PCC 7822 is an excellent cyanobacterial model organism with great potential to be applied as a biocatalyst for the production of high value compounds. Like other unicellular diazotrophic cyanobacterial species, it has a tightly regulated metabolism synchronized to the light-dark cycle. This strain was one of 6 Cyanothece sequenced by the DOE Joint Genome Initiative with the objective to learn more about the relationship of species within a genus. This strain has proven to be most valuable because it is transformable and a rudimentary genetic system has been established. It also has a large genome (~7 Mb) and is metabolically very robust. In order to use this strain as a biological chassis, an understanding of the diurnal metabolic dynamics is important. It is also critical to understand the phase and amplitude of mRNA transcripts and translated protein products relative to one another and the activity of genes located on the extrachromosomal genomic elements. Utilizing transcriptomic and proteomic methods, we were able to quantify the relationships between transcription and translation underlying central and secondary metabolism in response to nitrogen free, 12 hour light and 12 hour dark conditions.

Results: By combining iTRAQ based proteomics and RNA-sequencing transcriptomics, we quantitatively measured a total of 6766 mRNAs and 1322 proteins at four time points across a 24 hour light-dark cycle. Photosynthesis, nitrogen fixation, and carbon storage relevant genes were expressed during the preceding light or dark period, concurrent with measured nitrogenase activity in the late light period. We describe many instances of disparity in peak mRNA and protein abundances, and strong correlation of light dependent expression of both antisense and CRISPR-related gene expression. The proteins for nitrogenase and the pentose phosphate pathway were highest in the dark, whereas those for glycolysis and the TCA cycle were more prominent in the light. Interestingly, one copy of the psbA gene encoding the photosystem II (PSII) reaction center protein D1 (psbA4) was highly upregulated only in the dark. This protein likely cannot catalyze O2 evolution and so may be used by the cell to keep PSII intact during N2 fixation. The CRISPR elements were found exclusively at the ends of the large plasmid and we speculate that their presence is crucial to the maintenance of this plasmid.

Conclusions: This investigation of parallel transcriptional and translational activity within Cyanothece sp. PCC 7822 provided quantitative information on expression levels of metabolic pathways relevant to engineering efforts. The identification of expression patterns for both mRNA and protein affords a basis for improving biofuel production in this strain and for further genetic manipulations. Expression analysis of the genes encoded on the 6 plasmids provided insight into the possible acquisition and maintenance of some of these extra-chromosomal elements.

DNA Forensics in the Cloud

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.07

Seth Faith[1], Melissa Scheible[1]
[1]North Carolina State University, Forensic Sciences Institute

DNA forensics is on the brink of a revolution with new techniques in next-generation sequencing (NGS) and bioinformatics. Several recent studies have demonstrated NGS backward compatibility to existing criminal databases, plus expanded capabilities in higher genetic resolution (e.g., additional markers and sequence based-evidence) and also investigative leads from DNA (e.g., ancestry and physical appearance). One current barrier to implementation of NGS by forensic scientists is an effective analytical toolkit that can process the large and complex datasets produced from NGS instrumentation. Cloud computing may offer advantages for forensics laboratories to use NGS without tremendous capital investment and specialized staffing. Here we tested the potential of the Amazon Cloud computing environment for forensic microsatellite (STR) and mitochondrial genotyping. We first inspected the ease of data transfer between sequencer (Illumina MiSeq), cloud storage "buckets" (S3), compute clusters (EC2), and local devices (laptops). Data transfer was readily achieved via secured shell (SSH) and Amazon command line interface (CLI) tools. Next, varying Elastic Compute Clouds (EC2) instances were tested for optimized performance. For STR analysis, a previously developed algorithm, STRaitRazor, was deployed using a custom configured Linux Amazon Machine Image (AMI). Here, compute optimized instances were favored for cost and speed. For mitochondrial analysis, we tested a Windows Amazon Machine Instance with commercial off-the-self NGS analysis software. Using Microsoft remote desktop, secured connection could be established to conduct reference alignment and variant calling on a memory optimized AMI. Lastly, security measures were also evaluated. We found that two-factor authentication (security tokens and security groups) could be easily configured to allow for secured cluster computing. This study shows the base potential to perform low-cost, secured forensic NGS analysis using cloud-optimized methods. Additional analytical tool development and testing would be necessary for forensic end-users to adopt this technology in their laboratories, but this proof of concept demonstrates a path to NGS implementation without significant computational resource investment.

Interrogation of Interspecies Relationships in Algal Production Cultures

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.08

Juliette Ohan[1], Armand Dichosa[1], Momchilo Vuyisich[1], Pulak Nath[1], Shawn Starkenburg[1]
[1]Los Alamos National Laboratory

A growing body of literature suggests that many unicellular algae (microalgae) require bacteria to provide essential nutrients and metabolites for optimal growth and survival in natural microbial ecosysteThese same unicellular microalgae (e.g. Chlorella, Nannochloropsis) have been targeted for commercial applications because of their ability to efficiently accumulate biomass and/or lipids for conversion into renewable transportation fuels and other useful bioproducts. At best, there is anecdotal evidence regarding the extant bacterial mediated effects on algal growth or nutrient utilization. Research conducted to improve the productivity of commercial algal ponds have relied on the extrapolation of growth studies conducted under axenic laboratory conditions and have neglected to consider the positive (or negative) impact of bacteria. Dissecting the millions of possible algal-bacterial interactions is too great a task under current techniques. To discover and identify these interactions on a high throughput scale, we are developing a methodology to efficiently capture a small quantity of bacterial cells in porous agarose beads (gel microdroplets or 'GMDs') for cultivation with algae in a microfluidics chamber. We have developed a workflow to monitor growth within these GMDs and subsequently recover cells with the goal of identifying the bacteria that modify algal physiology to increase biomass yield and/or reduce nutrient input.. Thus far, we have optimized cell load input using fluorescence-activated cell sorting (FACS), recovered bacteria from our high throughput pipeline, and sequenced 16S rDNA of several species of interest. In parallel, we have developed a medium-throughput system using FACS that allows for the same analysis of cultures in a more traditional microbiological approach (not using microfluidics).

Herein we present progress made towards constructing this culturing system, monitoring bacterial loads in the test cultures, and demonstrating growth of algae within millions of picoliter sized environments. The impact of this work will positively influence the current commercial methods of algal biofuel production as growth-promoting bacteria are identified.

Validation of Whole Genome Average Nucleotide Identity for Identification of Listeria monocytogenes and related species

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.09

Lori Gladney[1], A. Huang[1], Z. Kucerova[1], L. Katz[1], K. Roache[1], H. Carleton[1], C. Tarr[1]
[1]Centers for Disease Control and Prevention

Listeria monocytogenes (Lm) is a Gram-positive bacterium and the only significant human pathogen in the genus Listeria. An infection with Lm is characterized by listeriosis, which may present as bacteremia and/or meningitis. Within Lm, there are four phylogenetically distinct lineages (LI – LIV) and most cases of disease are caused by strains within LI and LII. There are 16 additional Listeria species that have been described to date. Accurate and timely identification of this species is necessary; foodborne outbreaks caused by L. monocytogenes occur regularly and the fatality rate may approach 30% in immunocompromised individuals such as the elderly, cancer patients, and pre-term babies born with listeriosis.

Listeria species have traditionally been identified using a panel of phenotypic tests or molecular tests such as PCR (real-time or conventional) or with hybridization assays. With the advances in whole genome sequencing, all necessary information for identification and subtyping can be obtained from a single streamlined workflow that exploits the whole genome sequence (WGS). Average nucleotide identity (ANI) was first proposed to be a suitable method for circumscription of bacterial species in 2005 and offers improvements over traditional approaches. ANI can be calculated directly from the genome sequences, reducing the time spent on culture-based methods and other assays, thereby improving the timeframe for identification and subtyping lineages of L. monocytogenes.

Additional methods that utilize WGS have been proposed for identification of species and phylogroups such as rMLST and kmer-based approaches. While these methods may be reliable for species identification, there is significant time invested in building annotated databases and inference of phylogenetic trees, whereas the ANI can be computed quickly from high quality draft genome assemblies and is easily interpreted since it is a simple percentage.

We obtained whole genome sequences for 39 isolates of L. monocytogenes and 37 isolates of 15 other Listeria species, representing the diversity of the genus. The genomes were assembled with CG-Pipeline version 0.2. We made custom scripts to calculate ANI between two genomes using dnadiff in the MUMmer 3.0 package and ran all pairwise comparisons. We then performed a distribution analysis of ANI values in R. We sought to determine if the distribution of ANI values among the isolates showed a clear delineation of L. monocytogenes from the other species, and also if the clinically significant genetic lineages of Lm (LI and LII) were distinct as assessed by ANI.

The distribution analysis revealed that Lm can be clearly delineated from other Listeria species, and LI and LII were also distinct from each other and from LIII and LIV isolates based on the distribution of ANI values. Our data suggest that ANI is a suitable method for identification of Listeria species and can accurately identify the pathogen Lm. Thus, this method is an approach for Lm identification that can be incorporated into a streamlined WGS workflow. Moreover, this method may be used in real-time to rapidly identify bacterial pathogens.

Development and Evaluation of a Pan-viral Detection and Discovery Array

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.10

Clinton Paden[1], Ying Tao[1], Eishita Tyagi[2], Suxiang Ong[1]
[1]Division of Viral Diseases, Centers for Disease Control and Prevention, [2]Booz Allen Hamilton

Currently, efficient and rapid testing of clinical samples from unknown disease outbreaks is hindered by the lack of a comprehensive diagnostic system that has the capacity to sensitively detect known and unknown pathogens. We developed a panel of 38 sets of pan-viral group (family, subfamily, or genus) PCR primer sets which detect known and novel species of the virus families known or suspected to cause human disease. When there are many samples or time is critical, such as in an outbreak response, 38 individual PCRs plus sequence confirmation of each positive amplicon is inefficient. In this study we have validated and developed a streamlined system for efficiently testing multiple samples for a broad range of viral pathogens. We adapted the 38 generic PCR assays for use in a microfluidic device (Fluidigm Access Array), which automates setup and thermal cycling of up to 48 assays with 48 samples simultaneously (2304 reactions). The mix of amplicons is barcoded per sample and ready for sequencing on a desktop sequencer, such as the Illumina MiSeq. We developed a custom bioinformatics pipeline to process the resulting data from raw sequence to an easily-understandable report, which includes abundance estimates, alignments metrics, and amplicon consensus sequences for the identified taxa.

We evaluated the sensitivity of this system, compared to our traditional pan-viral family/genus PCR, which uses nested or hemi-nested RT-PCR/PCR reactions followed by capillary DNA sequencing. While the microfluidic PCR amplification uses much less sample volume per reaction (about 1/100th) compared to the traditional, nested PCR, the overall detection limit for 35% of the virus PCR reactions is comparable, and for 50% is within 1 log. We are currently evaluating preamplification/enriching strategies to ensure that these small volumes of sample contain enough target molecules to be distributed to all reactions for improved detection sensitivity.

Using the amplicons generated in this system, we are able to investigate dozens to hundreds of samples simultaneously on a desktop sequencer, which is not possible with shotgun metagenomic sequencing. This system may be used for identifying common pathogens from a large set of samples. It may be used for screening large sample sets for novel viruses, and interesting samples may then be chosen for full genome sequencing. These uses bridge the gap between the specificity, throughput, and low-cost of PCR and the power of next-generation sequencing.

Correlation of mutations detected in liquid and tissue biopsies

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.11

Eric Vincent[1], Douglas H. White[2], Douglas Horejsh[1], Molly A. Accola[3], William M. Rehrauer[3], Jeffrey Franz[1], Herly Karlen[1], Marjeta Urh[1]
[1]Promega Corporation, [2]Promega, [3]University of Wisconsin Hospital and Clinics

Circulating cell-free DNA (ccfDNA) in plasma can be used to detect biomarkers that show great promise for diagnosis and monitoring of cancer, giving rise to the possibility of liquid biopsies that obviate the need for invasive tissue collection. The low concentration and highly fragmented nature of ccfDNA, coupled with the low frequency of potential oncogenic biomarkers, present challenges that will require a purification method that is efficient and highly reproducible.

Here, we describe a method for purifying nucleic acids based on novel surface and binding chemistries. The combination of these two approaches allows for increased binding of fragmented DNA. The method can be partially automated to ensure highly reproducible results. Up to 4mls of plasma can be processed and eluted in 50ul, giving DNA concentrations of greater than 1ng/ul. It is possible to concentrate samples further with a subsequent purification step. This greatly facilitates use in Next Generation Sequencing.

We also describe an amplification-based measure of fragmentation for measuring ccfDNA quality. In a single qPCR well, amplifiable concentration, fragmentation, and inhibition can be assessed efficiently. Fragmentation is important to note as the small target is about 80bp and the larger target is about 294bp (much larger than the expected 170bp ccfDNA). Using this tool, genomic DNA contamination in the ccfDNA eluate can be seen.
Using this chemistry, ccfDNA was purified from the plasma of 7 patients who had previously undergone surgical resection for malignancy. DNA was also purified from the FFPE malignant tissue obtained from slides, following macrodissection, from the same patients. NGS was used to interrogate both sample types for potentially oncogenic variants. Several laboratory developed tests, all including COLD-PCR, were also employed to verify the presence or absence of variants. The two types of samples showed excellent correlation on mutations, suggesting that use of a less invasive liquid biopsy has the potential to enable actionable mutation detection without using more invasive solid tumor biopsy means.

Regulatory haplotypes in HLA-D coordinate transcription of the antigen presentation pathway

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.12

Edward Wakeland[1], Prithvi Raj[1], Ran Song[1]
[1]University of Texas Southwestern Medical Center, Dallas, TX 75390-9093

Systemic lupus erythematosus (SLE) is an autoimmune disease caused by a broad-based loss of humoral immune tolerance leading to the production of autoantibodies to a spectrum of self-antigens. Genetic predisposition is key for SLE susceptibility, however little is known about the nature or functional properties of causal genetic variants. We used targeted population sequencing to comprehensively characterize genetic variability at 28 risk loci for SLE in a panel of 1349 Caucasian SLE cases (773) and controls (576). The HLA-D region contains the strongest risk loci identified for SLE, with multiple alleles of both HLA-DR and -DQ showing strong associations. Sequence analysis of the 380 Kb segment spanning the BTNL2-DR-DQB2 region identified 15,261 common (MAF >0.05) genetic variants. Analyses of these sequence-defined HLA-D variations identified three independent risk-associated signals reaching genome wide significance. Subsequent analyses demonstrated that these disease-associated variations are imbedded in a series of stable haplotypes formed by multiple, ENCODE and eQTL-defined functional variations impacting the transcription of more than 20 genes that encode components of the antigen processing and presentation (APP) pathways of HLA class I and class II genes. Median neighbor joining analyses identified three HLA-D region regulatory haplotypes forming a risk clade strongly associated with SLE, all of which contained eQTL variants that increased the transcription of HLA-DR, DQ, DP, and other elements of the APP pathway in multiple myeloid and lymphoid cell lineages. This risk clade contains all of the classical HLA-D class II alleles previously associated with SLE, indicating that the systemic upregulation of the APP pathway is a consistent feature of all SLE-associated HLA-D alleles. Similar analyses of non-HLA SLE risk loci identified regulatory haplotypes that were often associated with transcriptional changes in multiple genes within specific pathways. Our analyses demonstrate that such regulatory haplotypes have increased disease-associated odds ratios in comparison to the disease odds for maximal GWAS tagging SNPs in these loci. These findings are consistent with the hypothesis that the functional variations that underlie many common disease alleles form regulatory haplotypes that modulate the transcription of multiple genes in immune system pathways and that their functional phenotypes are potent and complex.

CRPµTIC: Critical Reagents Program Microbial Threat Information Center- a Solution for Organization, Storage, Retrieval and Display of Microbial Data

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.13

Tyler Barrus[1], Kristin L. Jones[2], Danielle Montoya[1], Jane Tang[1], Mark J. Wolcott[3], Walter J. Berger[1], Michael A. Smith[2], Shanmuga Sozhamannan[2]

[1]Noblis, [2]Critical Reagents Program, [3]USAMRIID

Background: Increasingly sophisticated next-gen technologies have altered the manner in which bacterial and viral pathogens are characterized and have enabled improved phenotype-to-genotype linking. As such, genomes and phenomes can be generated at a rate that vastly outpaces the application of these datasets to medical countermeasure development. In order to manage the vast amount of data produced by such efforts, CRPµTIC has been developed. CRPµTIC is an information management system for the organization, storage, retrieval and display of microbial information from high-throughput phenotypic and genotypic characterization of biodefense pathogens and their near neighbors. This information is integrated with other pertinent strain information that can be used for product development.

Methods and Results: CRPµTIC is both a web-resident database and a low-common denominator process easily adapted by laboratories and enables the capture and central storage of data. CRPµTIC is based on open source software (e.g., PostgreSQL), is portable and has a full complement of strain and assay data administration functions to maintain data integrity. CRPµTIC captures key assay- and strain-related data in simple user interfaces, and appends the more detailed instrument-level source files for user download, with role-based access. The various data types include: sample metadata (organism ID, clinical, temporal, geographical and acquisition information); microbiological and biochemical characterization data; genomic data such as PCR, optical mapping, and whole genome sequencing; and finally data on various CRP products developed using these strains. The data are organized in a modular structure (Microbial Data Index (MDI), Microbial Metadata, Analytical Data, and CRP Products) through which the user can quickly navigate and identify records of interest by filtering on strain names and other identifiers (e.g., country of collection). Results can be sorted in various ways, including level of strain characterization. The unique feature of CRPµTIC is that all the information and products pertaining to a given organism are traceable to a single, well characterized microbial source stored in a repository, the Unified Culture Collection (UCC). UCC represents well-characterized microbes accessioned from geographically and temporally divergent sources that are subsequently characterized using a variety of technologies. The data organized in CRPµTIC can be output in the form of an MDI with hyperlinks to the specific analytical data pages that provide additional information on specific assay results. CRPµTIC currently contains over 1400 assay results, over 1300 organisms, over 1000 CRP products and has over 30 assay data types.

Conclusions: CRPµTIC serves as an end-to-end information management solution for microbes; i.e., from isolation of an organism all the way to product development. Although CRPµTIC does not offer analytical tools, the user can download the primary data and perform secondary analyses using their own analytical tools to answer specific queries of their interest. Also, the framework of CRPµTIC is easily adaptable to new assay types and is portable to different hosting platforCRPµTIC is intended to serve agencies across the USG as a reliable source of information about CRP reference materials and assays for the development of medical countermeasures including diagnostics and therapeutics.

CRISPR analysis of Yersinia pestis stains from Georgia

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.14

Ekaterine Zhgenti[1], Mari Murtskhvaladze[1], Anna MacHablishvili[1], Gvantsa Chanturia[1],
Tea Tevdoradze[1], Tracy Erkkila[2], Patrick Chain[2]
[1]National Center for Disease Control and Public Health/Lugar Centre, Tbilisi,
[2]Los Alamos National Laboratory

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) are particular family of tandem repeats found in a wide range of prokaryotic genomes. They are composed of highly conserved short repeated sequences (DR) 21–37 bp in length, interspaced with unique non-repetitive elements or “spacers” and associated with genes involved in DNA recombination and repair. In the Yersinia pestis genome three such elements YPa, YPb and YPc are found at three distinct loci and presently 137 spacer sequences have been reported. The distribution of spacers and their arrays in Y. pestis strains is strongly region and focus specific and can provide important information for genotyping and evolutionary research of bacteria, which could help to trace the source of outbreaks.

The purpose of our study was CRISPR-based analysis of three Y. pestis strains isolated from two natural plague foci of Georgia; two strains from high mountainous focus (Transcaucasian highland, bordered with Armenia) and one from the plain foothills (encompassing the eastern part of Georgia, next to Azerbaijan border).
The whole genome sequencing (WGS) of the isolates were performed using a 2x300bp sequencing chemistry on the next-generation sequencing (NGS) Illumina MiSeq Platform. WGS data was further analyzed with the CLC Genomics Workbench software package (CLC bio) EDGE Bioinformatics.

The spacers arrays were acquired and analyzed online using the ‘‘CRISPR Finder Tool’’ and ‘‘spacers dictionary’’ tools in CRISPRs database (http://crispr.u-psud.fr/). DRs and spacers were identified and extracted. The selected spacer sequences were compared against the microbial genome database in GenBank (http://www.ncbi.nlm.nih.gov).

The present study is the first attempt of CRISPR-based analysis of Y. pestis strains in Georgia. The data are still under processing and results will be presented in poster presentation.

A Hybrid Graph Approach for Short-Read Assembly

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.15

Philipp Kämpfer[1]
[1]Heidelberg Institute for Theoretical Studies

Since the advent of DNA sequencing methods in the late seventies a variety of new technologies have been developed allowing for cheaper and faster DNA sequencing. Though many consider the assembly of DNA fragments a solved computational problem, unordered and fragmented genome assemblies with false joins are widespread, significantly hampering any downstream analysis. Next-generation sequencing (NGS) methods have proven to be low-cost and high-throughput through the parallelization of the sequencing process, however at the cost of short read-lengths. The short read lengths and the additional shortening by decomposing them into k-mers to efficiently build a de Bruijn graph are the primary reasons why, despite high sequencing coverage, most assembly tools have difficulty producing accurate assemblies with long-range contiguity.

Currently, the reconstruction of genomic information using short-read data utilizes two distinct graph-based approaches, namely the overlap-layout-consensus concept (OLC) and de Bruijn graphs (dBG). The overlap-layout-consensus (OLC) concept is considered superior to the de Bruijn graph (dBG) in that the unit of assembly is a read as opposed to a small k-mer, causing the graph and its path structure to be simpler and easier to disambiguate, together resulting in higher contigs lengths. However, popular NGS assemblers rely solely on the de Bruijn graph due to its superior runtime efficiency compared to the quadratic nature of the OLC approach.

We present a new hybrid graph approach, which includes a variety of novel graph algorithms for fast and effective error correction. These eliminate more than 99% of all sequencing errors in linear time while taking advantage of both approaches. We combine the fast, linear-time construction of a dBG with the higher contig resolution of the OLC approach. This is accomplished by touring the dBG and collecting read information while simultaneously constructing an overlap graph directly from an expanded dBG, containing information on the original sequencing reads prior their decomposition into k-mers. By further deriving a string graph from the transitively reduced overlap graph we are able to reconstruct large unique contigs of the genome. Finally, the paired-end read information is incorporated into the string graph to facilitate scaffolding-like contig ordering and resolution of repetitive sequences. The hybrid graph approach can be used with various insert sizes and sequencing technologies.

Indeed, tests on several bacterial short read data sets have shown that the hybrid approach is in time and space complexity comparable with state of the art dBG assemblers like velvet but provides the higher contig resolution of an established OLC assemblers as SGA (string graph assembler) avoiding the long run-time and storage consumption. On this account the hybrid graph concept and the developed algorithms have a high potential for a further improvement of the quality of the denovo genome assembly based on short reads.

An improved assembly algorithm for de novo circular genome reconstruction

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.16

***Christian Olsen*[1], *Kashef Qaadri*[1], *Helen Shearman*[1], *Richard Moir*[1], *Matt Kearse*[1], *Simon Buxton*[1], *Jonas Kuhn*[1], *Matthew Cheung*[1]**

[1]*Biomatters*

Circular chromosomes or genomes, such as viruses, bacteria, mitochondria and plasmids, are a common occurrence in nature, but despite the wide array of algorithms available for de novo assembly, the circularity of these DNA molecules is largely overlooked. There are some limitations of this oversight. Current NGS assembly algorithms assume a linear molecule and will result in a linearly represented genome, with a breakpoint in an arbitrary position. This is becoming increasingly problematic with increasing NGS read lengths resulting in fewer contig assemblies with less ambiguity. Additionally, long read sequencing technologies promise to provide sequences that may greatly span breakpoints at the expense of coverage - resulting in a relatively large quanta of information loss.

Although there are currently methods for the re-circularisation of contigs post-assembly by identifying common trailing/leading sequence motifs, we present a more robust approach of circularising during the assembly process whilst still allowing the merging of similar and sub-contigs throughout the overlap-based approach. This method can also combat the issue of chimeric sequence due to contamination, a common problem when sequencing bacterial cultures, due to decreased likelihood of conserved contig ends due to timely circularisation.

We present results from both 28.6 million read Pan troglodytes illumina data set and 267,491 read Panthera leo persica (Asiatic Lion) mitochondrial NGS library produced using an Ion Torrent sequencing machine. These results are discussed and compared to some of the more popular linear assembly algorithms in common usage today.

Geneious R8 is the first bioinformatics software package to offer a circular de novo assembly method. The Geneious Circular de novo assembler is developed by Biomatters and may be found at http://www.geneious.com

Longer, More Accurate Sequences on the Ion Torrent Platform

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.17

***Sihong Chen**[1], **Eileen Tozer**[1], **Mindy Landes**[1], **Theo Nikiforov**[1], **Anelia Kraltcheva**[1], **Guobin Luo**[1], **Kevin Heinemann**[1], **Josh Shirley**[1], **Peter Vander Horn**[1], **Daniel Mazur**[1]*

*[1]**Thermo Fisher Scientific***

Ion Torrent's chip-based sequencing technology is well suited to quickly, cheaply, and now (with Hi-QTM) accurately sequence samples of all types. The Ion PGMTM Hi-QTM Sequencing kit has decreased overall error by 50% and decreased systematic (non-random) error by up to 90%. Use of Hi-QTM Sequencing produced substantial improvements in accuracy, systematic error, and readlength on templates ranging from 100 to 400bp. Moreover, we will soon extend our readlength capabilities into the 600-800bp range. Such improvements in the PGM system enable a broader range of applications, such as enhanced de novo genome assemblies, Human Leukocyte Antigen (HLA) sequencing, bacterial identification, and meta-genomic analysis.

Genome sequence of a North American Borrelia miyamotoi isolate using PacBio long reads

Wednesday, 27th May 18.30 - Mezzanine (Sponsored by Roche) – Poster 2a.18

Luke Kingry[1], Adam Replogle[1], Dhwani Batra[2], Martin Williams[1], Marc Dolan[1], Jeannine Petersen[1], Martin Schriefer[1]
[1]Centers for Disease Control and Prevention, Fort Collins, CO, [2]Centers for Disease Control and Prevention, Atlanta, GA

The relapsing fever spirochete, Borrelia miyamotoi, is found widely distributed in Ixodes spp. ticks in Europe, Japan, and North America. The public health impact of B. miyamotoi has been highlighted by recent reports of human disease caused by the bacterium from Europe, Russia, Japan, and the United States. The Borrelia genome is comprised of a 1 Mb linear chromosome and several circular and linear plasmids of varying size that encode highly redundant sequence making accurate plasmid assemblies challenging. To date there is no publicly available plasmid sequence for B. miyamotoi. To circumvent assembly problems with repetitive plasmid DNA, whole genome sequencing was conducted using the Pacific Biosciences RS II instrument and resulted in an average read length of 7 kbp. Sequencing analysis and contig assemblies were conducted using the SmrtAnalysis 2.2.0 software suite and the hierarchical genome-assembly process (HGAP). The assembly generated a single 907,260 bp linear contig representing the main chromosome and 3 circular and 7 linear plasmids from 19-72 kbp. The chromosome was predicted to encode 889 open reading frames, 32 tRNAs, and 3 rRNAs. Plasmids were found to encode highly redundant arrays of variable large and small proteins (vlp/vsp) for antigenic phase variation. The ability to generate near complete genome assemblies of complex Borrelia genomes with long reads will allow for a more complete understanding of the highly redundant plasmid DNA found in both relapsing fever and Lyme disease causing Borrelia.

Repair of challenging FFPE DNA improves library success rate and sequencing quality

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.01

***Fiona Stewart**[1]**, Pingfang Liu**[1]**, Lixin Chen**[1]**, Laurence Ettwiller**[1]**, Richard Corbett**[2]**, Helen McDonald**[2]**, Pawan Pandoh**[2]**, Christine Sumner**[1]**, Eileen Dimalanta**[1]**, Theodore Davis**[1]**, Yongjun Zhao**[2]**, Marco Marra**[2]**, Thomas Evans**[1]*

*[1]**New England Biolabs, Inc.,** [2]**Genome Sciences Centre, BC Cancer Agency***

Formalin-fixed, paraffin-embedded (FFPE) clinical samples are an invaluable source of information about genetic alterations in human disease, especially cancer. Next generation sequencing is a powerful tool to mine that information. Unfortunately, sequencing DNA from FFPE samples is challenging due to limited quantities and poor quality, a result of DNA damage incurred during fixation and storage.

In this study, we investigated the effects of DNA repair on library preparation and sequencing from FFPE samples. We evaluated a repair enzyme mix that is designed to work on a broad range of DNA damage including modified bases, nicks, gaps, and blocked 3' ends. Results from FFPE DNA samples of varying quality show that DNA repair generally increases library yields, and improvements from 10% to 458% have been observed. Careful analysis of sequencing data shows that base calling qualities for all 4 bases are improved upon DNA repair. Aberrant G:C to A:T mutation was significantly reduced, consistent with cytosine deamination being induced during fixation and storage at room temperature. Interestingly, sequencing miscalls for base pair changes not typically associated with fixation were also reduced. With DNA repair, a noticeably positive impact on read mapping and read pairing was observed, resulting in the generation of more useable data. Pretreatment of FFPE DNA by the repair enzyme mix is easily implemented upstream of library construction, allowing fast turnaround time and easy automation. We expect that these improvements will enable the analysis of many FFPE samples that would otherwise not be accessible.

Understanding genome evolution in non-model taxa is negatively affected by homology based, transposable element identification

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.02

***Roy Platt*[1], *David Ray*[1]**

[1]*Texas Tech Univeristy*

Transposable elements (TEs) occupy large portions of eukaryotic genomes, are highly variable, and may have significant impacts on the biology and evolution of organis Therefore a proper and thorough annotation of newly sequenced genomes is of utmost importance. Often the repetitive portion of the genome is ignored in favor of protein coding genes or their regulators. In most cases TEs are identified based on homology to known TEs from related taxa. Unfortunately for these projects, TE annotations via homology are, by definition, only able to identify TEs homologous to known elements. The result is that lineage specific subfamilies or even entire classes of TEs deposited via horizontal transfer may be missed giving an inaccurate picture of the TE landscape. To demonstrate this, we preformed de novo TE analyses in two rodent genomes, the prairie vole (Microtus ochrogaster) and the naked mole rat (Heterocephalus glaber). Age distributions and overall TE content identified in each species varied based on distance to Mus musculus and Rattus norvegicus. In M. ochrogaster, 27.8% of the genome was derived from TEs with more than 110 Mb of lineage specific L1 identified through de novo analyses. Similarly, an additional 90 Mb of TEs were identified in the H. glaber genome overlapping 380 exonic regions, representing potential exaptation events. Additional analyses across all available mammal genomes demonstrate that TE annotation quality varies based on distance to the closest “model” species. These observations demonstrate the necessity for de novo TE annotations in order to understand the activity of lineage specific TEs across mammal genomes.

Assembly Strategy for the BAC Pool Sequencing of Aegilops tauschii, the Ancestor of the Wheat D Genome

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.03

Daniela Puiu[1]
[1]***Johns Hopkins University***

Previous efforts to sequence wheat genomes using the whole-genome shotgun (WGS) method have resulted in fragmented assemblies, much smaller than the estimated genome size. This is likely due to polyploid and unusually high repetitive content of the wheat genome.

We sequenced the 4.3Gb Ae. tauschii genome using 5,607 pools of 6 to 10 overlapping BACs. Each BAC is approximately 145Kb in length, while the pools span ~1Mb. The pools were sequenced using over 100x coverage of Illumina MiSeq paired-end reads. In addition ~12x coverage of public WGS mate-pair reads and ~7x coverage of PacBio reads were used for scaffolding, with plans to increase the PacBio coverage up to ~50x.

SOAPdenovo2 was used to generate the original BAC pool assemblies. ABYSS konnector was used for MiSeq read extension while SSPACE-LongRead was for PacBio guided scaffolding. We also implemented new methods for long terminal repeat assembly and WGS mate-pair/PacBio pool assignment. Using these methods we were able to increase the scaffold size from from an original of ~100Kb to over 500Kb.

We will present these methods along with a new version of AMOS minimus assembler which is being used for merging the pools.

This work is a part of NSF-funded Project IOS-1238231 to generate a reference sequence for the genome of Ae. tauschii (http://aegilops.wheat.ucdavis.edu/ATGSP/).

Error correction of Illumina MiSeq® reads as a critical step in genome assembly and SNP analysis

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.04

Darlene Wagner[1], Heather Carleton[1], Eija Trees[1]
[1]Enteric Disease Laboratory Branch, Centers for Disease Control and Prevention

In next-generation sequencing (NGS), low-quality base calls may degrade genome assembly, single-nucleotide polymorphism (SNP) discovery, or other downstream analyses. It is demonstrated here that QUAKE (Kelly et al., 2010), a kmer-based sequence error correction tool, improves assemblies and SNP discovery for clinical isolates of Salmonella enterica and Shiga-toxin-producing E. coli (STEC). A set of eight Salmonella serovar Enteriditis Illumina MiSeq reads from laboratory A exhibited pairwise SNP differences ranging from 32 to 171 (median 102), despite epidemiologic evidence of origins from a single-source outbreak. In addition, laboratory A data exhibited discrepancies in quality scores (Q30) between the forward (R1) and reverse (R2) reads, averaging 36.3 and 33.4, respectively. After correcting laboratory A data through QUAKE, quality scores of both R1 and R2 reads averaged 36.0. Corrected reads then yielded less-fragmented assemblies in CLC Genomics Workbench (Qiagen, Aarhus, Denmark) relative to the uncorrected reads. Median number of contigs went down from 51.5 to 42 while average N50 increased from 268,804 to 321,129. After QUAKE trimming, pairwise SNP differences among the eight Enteriditis strains ranged from 6 to 21 (median = 14), comparable to laboratory B Illumina MiSeq reads of the same outbreak cluster (2 to 14, median 8.5). This indicates QUAKE correction minimizes phylogenetically-irrelevant SNP calls between related strains. Eight STEC representing O45 and O145 serogroups similarly exhibited quality score discrepancies between R1 and R2 reads, averaging 35.6 and 31.4, respectively. Correction through QUAKE yielded quality scores of 35.1 for both forward and reverse reads. QUAKE correction improved non-scaffolded CLC Genomics Workbench assemblies, decreasing median number of contigs from 227 to 209 and increasing N50 from 110,562 bp to 120,630 bp. The median SNPs between O145 and O45 reads increased from 11,879 to 12,299, suggesting QUAKE aids discovery of phylogenetically-relevant SNPs. Given that there is inevitable sequence quality variation that may stem from mechanical issues, technician experience, and proximity to reagent expiration date, k-mer-based correction should be considered for bioinformatics pipelines incorporating Illumina reads from multiple laboratories.

Highly efficient sequencing of viral genomes in complex samples

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.05

Cheryl Gleasner[1], Omar Ishak[1], Andrew Hatch[1], Kim McMurry[1], Cheriece Margiotta[1], Tracy Erkkila[1], Jennifer Harris[1], Momchilo Vuyisich[1]
[1]Los Alamos National Laboratory

Sequencing full genomes of RNA viruses in clinical samples and culture supernatants is very important for detection and tracking of outbreaks, epidemiology, and understanding of genotype-to-phenotype relationships. Due to the very low relative abundances of viral RNAs in such samples, meta-transcriptomic sequencing requires generation and analysis of massive amounts of data.

We have developed an efficient and inexpensive virus sequencing pipeline using ultracentrifugation for viral enrichment and ultra-sensitive directional RNA library preparation methods. We have applied these methods to sequencing human Influenza A virus genome in human serum, plasma, and cerebrospinal fluid (CSF), and in the culture supernatants. Ultracentrifugation enables up to 1000 fold enrichment of viral reads, and our library preparation method can utilize as little as 100 pg of input RNA. Combined with advanced read mapping and genome assembly tools, our laboratory methods enable small genome centers to respond to large outbreaks and support numerous epidemiology studies.

The Value of Improved Microbial Genome Assemblies

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.06

***Karen Davenport**[1], **Shannon Johnson**[1], **Hajnalka Daligault**[1], **Tracy Erkkila**[1], **David Bruce**[1]*
*[1]**Los Alamos National Laboratory***

A long standing debate in the microbial genomics community has focused on the value of bringing genomes to finished quality versus completing a greater number of genomes to a lower quality standard. The argument for quality assemblies lies in the need for confidence in the gene and gene cassette identification to determine functionality, particularly for antibiotic resistance, toxin production, assay specificity and accuracy. While comparisons of assembly type and product usage have been made before, none have been published recently without trying to suggest a single assembly or pipeline. Our goal is to determine which of the current assembly methods enables the most useful functional annotation of a pathogen genome.

We compare the differences of various levels and combinations of genome assemblies on functional annotation for the four genera: Bacillus, Burkholderia, Francisella and Yersinia. Our study will look at the statistics, after assembly and annotation, for the following genome assemblers: Velvet, IDBA, AllPaths and HGAP as compared to a finished reference quality genome assembly for each. To determine quality differences we will compare gaps, SNPs/INDELS, gene count, presence/absence of known pathogenicity islands/plasmids, virulence markers and rapid phylogenetic placements. Preliminary finished genome comparisons suggest that HGAP misses small plasmids, and CRISPRS and other pathogenicity markers located in tandem repeats are not resolved without HGAP. We will present detailed comparisons of assembly and annotation statistics along with preliminary indications as to the efficacy of phylogenetic placement using either partial or finished genome data.

Blended assemblies with manual review are still the only method to ensure that data is complete and accurate for downstream analyses. Often phylogenetic placement using housekeeping genes is correct using high quality Illumina std data but functional analyses regarding pathogenicity, resistance or other physiological factors cannot be determined. Inaccurate or incomplete information regarding the capability of a pathogen may delay proper identification and/or diagnosis.

Integration of single molecule, genome mapping data in a web-based genome browser for evaluating sequence based structural artifacts

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.07

William Chow[1], Matthew Dunn[1], Jonathan Wood[1], Kerstin Howe[1]
[1]Wellcome Trust Sanger Institute

The Genome Reference Consortium's (GRC) continued role in releasing improved reference assemblies for human (GRCh38), mouse (GRCm38) and zebrafish (GRCz10) drives an on-going need to modernize the underlying genome curation by consulting all relevant supporting data, including novel data types for emerging technologies.

In order to aid this curation process, we have developed the genome evaluation browser, gEVAL (geval.sanger.ac.uk). The gEVAL browser provides a one-stop solution to assess the compliance of a given path through a multitude of available data such as the correct pairing and suitable distance of mapped clone ends, the placement of markers and cDNAs, overlap evaluation, self-comparisons, multi-assembly comparisons, and many more.

Genome maps produced by single molecule optical mapping (OM) technologies such as those provided by Bionano Genomics and Opgen, are novel data types that have been recently integrated into the browser. Along with the wide range of aligned data already viewable on each genome, OM data can help identify and confirm assembly irregularities such as insertions, deletions and mis-assemblies while providing suitable information to resolve them.

The long-range nature of the OM data also provides scaffolding information for yet unplaced assembly components. While much of the reference genomes maintained by the GRC are already well represented at the chromosome level, confirmation of contig or scaffold component order remains valuable.

gEVAL updates are released frequently in between major GRC assembly releases, allowing the user an up-to-date snapshot of the evolving assembly, which may not be represented in other public databases.

After proving helpful with the curation of the GRC reference genomes, gEVAL
browser has been extended to include assemblies from rat, pig, chicken and several helminth organisms.
The majority of the curation and genome annotation data gathered in gEVAL is also available as a trackhub (http://ngs.sanger.ac.uk/production/grit/track_hub/hub.txt) to aid accessibility in browsers of choice.

Blending art and science to understand genotype and phenotype with public data and Molecule World™ on the iPad

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.08

***Todd Smith*[1], Sandra Porter[1]**
[1]*Digital World Biology*

It is impossible to grasp fundamental concepts in biology without understanding the relationship between sequence, structure, and function. Modern data collection technologies are creating enormous data resources that can be used to help students' understand these relationships, but currently are underutilized. This is due to the fact that easy-to-use tools that meet teachers' needs for clear instruction and data visualization are not yet commonplace. Filling the gap between the embarrassment of data riches and practical classroom use requires three things: user-friendly tools, content that demonstrates specific applications with interesting stories, and packages that combine instruction, assessments, and inquiry-based investigations.

Digital World Biology is addressing this need with its on-line courses and mobile apps. The on-line courses increase students' computer literacy while using standard tools like Cn3D, Blast, ORF finder, and multiple databases, in directed and exploratory ways, and help students better understand biology as well gain a better appreciation for the value of the data and the field of bioinformatics. In response to nearly two hundred interviews with K-12 and college teachers and students, we created Molecule World and the Molecule World DNA Binding Lab™ iPad apps to display 3D-data from multiple structure databases (MMDB, PDB, and PubChem) using a novel rendering engine that allows us to uniquely highlight chemical properties, sequence orientation, and a molecule's biochemically important features. The ability to display and highlight sequences and specific components within molecular complexes enables exploration into the relationships between sequence, structure, and function in new ways. Preliminary data collected in professional development workshops, and many demonstrations, supports the hypothesis that being able to view and simultaneously interact with data improves teaching capabilities and student engagement while making fun works of art!

Work supported by NSF grant IIP 1315426

Molecular characterization of the role of RUNX1 in Notch signaling in T-cell Acute Lymphoblastic Leukemia (T-ALL)

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.09

Rashedul Islam[1], Catherine Jenkins[2], Luolan Li[3], Alireza Lorzadeh[3], Misha Bilenky[3], Annaick Carles[3], Vincenzo Giambra[3], Sonya Lam[2], Catherine Hoofd[2], Miriam Belmonte[2], Xuehai Wang[2], Andrew Weng[4], Martin Hirst[5]
[1]University of British Columbia, Vancouver, [2]BC Cancer Agency Research Centre,
[3]Canada's Michael Smith Genome Sciences Centre, [4]University of British Columbia,
[5]Centre for High-Throughput Biology, University of British Columbia

T-cell acute lymphoblastic leukemia (T-ALL) is a hematopoietic malignancy driven by oncogenic activation of NOTCH1 signaling. T-ALL accounts for 15% of pediatric and 25% of adult acute lymphoblastic leukemia cases (Pui et al, N Engl J Med. 2004;350:1535–48). Previous studies have suggested that Notch1 signalling activity is increased in T-ALL patients (Weng et al, Science 306:269;2004). Additionally, NOTCH1 is capable of “evicting” Polycomb Repressive Complex 2 (PRC2) from target loci in T-ALL (Ntziachroistos et al, Nat Med 18:298;2012). Other work suggests that NOTCH1 may only load onto chromatin after modification by runt-related transcription factor 1 (RUNX1), a so-called “pioneer factor” (Terriente-Felix et al, Development 140:926;2013).

Based on this evidence we hypothesize that RUNX1 is required for oncogenic Notch signalling in T-cell leukemia. In support of this hypothesis we find that human T-ALL patient-derived samples and cell lines are sensitive to both RUNX1 depletion by lentiviral shRNAs and pharmacologic Notch pathway inhibition. In order to dissect the molecular mechanism(s) underlying these phenotypes, we have performed ChIP-seq against a panel of histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K27me3 and H3K9me3) for samples that have either been depleted of RUNX1 (shRNA) or NOTCH1 (pharmacologic inhibition) and identified enriched regions using FindER (http://www.epigenomes.ca/finder.html). From these data we have identified 12,487 and 19,190 enhancer regions are unique to the Notch-on and Notch-off conditions, respectively. GREAT functional gene set enrichment prediction (GREAT.stanford.edu) showed T/B cell activation related terms in the Notch-on specific enhancers, but not in the Notch-off sample. These Notch dependent enhancer elements are expected to be enriched in Notch binding sites that are lost when Notch is turned off. We also found an enhancer of 1,458 bp length, located 29,024 bp downstream of the Notch1 gene at chromosome 9 that is unique in the “Notch on” sample. Our study suggests that inhibition of NOTCH1 signalling results in a restructuring at the enhancer landscape potentially leading to loss of T/B cell activation. Our ongoing work aims to 1) prioritize RUNX1/NOTCH1 candidate loci by expression and further epigenetic criteria and 2) determine what Notch-independent effects RUNX1 has on the regulatory landscape.

An automated and efficient computational pipeline for the design of real-time PCR assays for diverse pathogen detection and biosurveillance

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.10

Norman Doggett[1], Murray Wolinsky[1], Jason Gans[1]
[1]Los Alamos National Laboratory

The challenges of designing sensitive, pathogen-specific PCR assays depend on the genetic composition of the target pathogens and their non-target near neighbors. Unique signature oligonucleotides (oligos) for bacterial pathogens at the species level are readily found using traditional sequence comparison methods (like BLAST). These signature oligos can form the basis for PCR assays by providing unique primer or probe binding site(s) that insure target specificity. However, this traditional assay design approach can fail when tasked with differentiating bacterial pathogens at the strain or subspecies level (e.g. Brucella suis vs other Brucella species) that differ primarily by genomic rearrangements and deletions. In addition, highly diverse targets (e.g., RNA viruses) traditionally require the use of computationally expensive multiple sequence alignments to identify the most highly conserved regions of these viruses, as well as the use of degenerate bases in primer and probe regions. Over the past 10 years, we have developed a suite of computational algorithms for solving these challenging nucleic acid-based assay design problems.

SPAdes family of tools for genome assembly and analysis

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.11

Dmitry Antipov[1], Anton Bankevich[1], Elena Bushmanova[1], Alexey Gurevich[1], Anton Korobeynikov[1], Sergey Nurk[1], Andrey Prjibelski[1], Yana Safonova[1], Alla Lapidus[1], Pavel Pevzner[2]

[1]Saint Petersburg State University, [2]University of California San Diego

Currently discovering a new fusion gene known for triggering cancer in patients presents significant experimental and computational challenges. For example, RNA-Seq data has high variations in coverage depth due to the different expression levels of various isoforms and genes, which further complicates its analysis. It could be said that RNA-Seq data shares many of the same computational complications as the single-cell MDA amplified data. This means that many of the algorithms developed for handling single-cell data can be reused for transcriptome assembly. We are presenting rnaSPAdes – a novel transcriptome assembler built on the top of the SPAdes platform.

The ability to generate large sets of RNA-Seq data created a demand for both de novo and reference-based transcriptome assemblers. Despite the fact that a number of manuscripts describing novel assemblers have been published, none of the studies focus on the comparison and benchmarking of the different tools. rnaQUAST is the first tool created for the purpose of evaluating RNA-Seq assembly quality and benchmarking transcriptome assemblers using reference genome and gene annotation. rnaQUAST calculates various metrics that demonstrate completeness and correctness of assembled transcripts, and outputs them in a user-friendly summary report.

Various companies are currently working on developing such long reads sequencing technologies as for example the TruSeq Synthetic Long Reads technology recently presented by Illumina. The unique feature of this technology is that it generates virtual long reads that are not the direct output of a sequencing machine, but a result of assembly of barcoded pools of short reads. It is the first approach to generate accurate long reads (less than 0.1% error rate) providing an attractive alternative to the technologies developed by Pacific Biosciences and the Oxford Nanopore (more than 15% error rates) with respect to both accuracy and cost. We are introducing truSPAdes - the first publicly available assembler specifically designed to cope with the complications of pooled barcode data.

Reference Assembly Creation for the Mouse Genomes Project

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.12

Jonathan Wood[1], Kerstin Howe[1]
[1]Wellcome Trust Sanger Institute

The Mouse Genomes Project is a collaborative effort to sequence the genomes of 18 key laboratory mouse strains and species widely used by the scientific community. The project has already provided accurate SNP and structural variant information [1].

Initial draft genome assemblies using data from multiple illumina short read insert libraries have been created using SOAPdenovo, improved via scaffolding and separated into pseudochromosomes through alignment to the GRCm38 reference. From this starting point in an effort to provide reference quality assemblies for the strains, the Genome Reference Informatics Team will manually improve each genome assembly using the Genome Reference Consortiums pipeline and additional data sources of optical mapping and Pacbio long read technologies. This will allow for a robust genome wide comparative study and the creation of a comprehensive and centralised repository with the information.

We will present an overview and the aims of the project for an initial comparative study of chromosome 11 using high quality manual annotation.

[1] Mouse genomic variation and its effect on phenotypes and gene regulation. Keane et al 2011 Nature 477, 289–294

A combination of genetic and genomic approaches to unravel the complexity of bacterial evolution

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.13

***Dongping Wang**[1], **Robert Dorosky**[2], **Cliff Han**[1], **Armand Dichosa**[1], **Patrick Chain**[1], **Jun Myoung Yu**[2], **Elizabeth Pierson**[2], **Leland Pierson**[2]*

[1]Los Alamos National Laboratory, [2]TAMU

Bacterial adaptation to stress condition is a complex process involving changes in gene content and expression. In this study, we use a combination of genetic and genomic approaches to study the evolution of Pseudomonas chlororaphis 30-84 small colony variant (SCV). The SCV demonstrated pleiotropic phenotypes including small cell size, slow growth and motility, but increased biofilm formation and resistance to antimicrobials. To better understand the genetic alterations underlying these phenotypes, RNA and whole-genome sequencing analyses were conducted by comparing a SCV mutant to the wild-type strain. Of the genome's 5,971 genes, transcriptomic profiling indicated that 1,098 (18.4%) have undergone a significant reprograming of gene expression in the SCV mutant. Whole-genome sequence analysis revealed multiple alterations in the SCV, including mutations in yfiR (cyclic-di-GMP production), fusA (elongation factor), and cyoE (heme synthesis) and a 70-kb deletion. Genetic analysis revealed that the yfiR locus plays a major role in controlling SCV phenotypes, including colony size, growth, motility, and biofilm formation. Moreover, a point mutation in the fusA gene contributed to kanamycin resistance. Our data also support the idea that phenotypic switching in P. chlororaphis is not due to simple genetic reversions but may involve multiple secondary mutations. These results highlight the strength of coupled genetic and transcriptomic analyses, in combination with whole-genome sequencing, as an approach to unravel complex bacterial adaptations.

Multi-omics computational pipeline for systems biology

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.14

Seongwon Kim[1]
[1]National Research Council

We aim at building computational pipeline that would be of generic use for systems biology research. Recent advancement of computational and algorithmic tools have enabled researchers to produce and manage high-throughput data at low costs, opening an unprecedented possibilities for analysis of microbial communities as a whole from the environment. Computational analysis of the sequence data obtained from next-generation sequencing often involves clearly delineated downstream workflow, where each step requires optimized and flexible utilization of various algorithmic tools. The pipeline is designed to integrate DNA, RNA, protein and metabolome dataset obtained from the environmental samples.

Typically, metagenome sequences obtained from whole-genome shotgun sequencing will undergo quality filtering, assembly, coverage calculation, taxonomy assignment and abundance profiling. Marker sequences such as 16S rRNA sequences can produce more refined taxonomy and phylogeny assignment using dedicated database and toolsets. Powerful parallel computing tools have enabled database search for the entire raw sequences at higher speed and low costs. Analysis of RNA data (metatranscriptomics) enables researchers to decipher gene expression of the community under various conditions, shedding light on the community response and strategy for optimal resource exploitation. Proteomics methodology such as liquid chromatography and tandem mass spectrometry offers invaluable opportunity for the validation and illumination on the community functional profiles. Metabolomics component of the analysis concerns the reactants and products of the enzymatic reactions of the community. Metagenomic sequence data can produce multi-faceted prediction and assessment of such metatranscriptomics, metaproteomics and metabolomics characterization of the communities, where typical analysis steps would include gene prediction, database search, mapping into reaction network and experimental validation.

The pipeline (called MOCA, Multi-Omics Computing Aid) facilitates the processing, corroboration and visualization of such datasets in a user-friendly manner. The pipeline aims at being an automatable, modular and robust toolset to be of generic use for meta-omics data for systems biology. It is expected to provide researchers with tools that can receive flexible input, and perform data processing and analysis with minimum intervention, enabling standardized assessment and comparison of the results for many different samples and workflows. Given the vast and increasing number of tools, the pipeline is designed to be modular, where new tools can be easily integrated and manipulated by users. Different data formats at different levels of the pipeline are treated in a compatible manner, while offering options for varied methodology and evaluation. The pipeline is expected to offer a coherent, robust and flexible toolset for incorporating complex arena of systems biology.

Evaluation of HiSeqX Ten Performance in a Production Pipeline

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.15

Kimberly Walker[1], Christian Buhay[1], Rashesh Sanghvi[1], Qiaoyan Wang[1], Harsha Doddapaneni[1], Jianhong Hu[1], Mark Wang[1], Yi Han[1], Huyen Dinh[1], Eric Boerwinkle[1], Donna Muzny[1], Richard Gibbs[1]

[1]Baylor College of Medicine

Implementation of the HiSeqX Ten fleet has been a major technical focus for the HGSC since the fall of 2014. To date, six instruments have been delivered resulting in the completion of over 47 flow cells and 375 30X genomes. As with any new platform scheduled for production implementation, a complete examination and optimization of analysis pipeline processes is required along with development of new quality metrics to assure optimal pipeline performance.

At the start of our implementation, we sought to learn the typical characteristics of samples from the HiSeqX Ten. Initially, we noticed unique characteristics not seen in other Illumina platfor We observed four times more duplicates per sample in comparison to capture-based exomes. We detected an increase in clustered optical duplicates correlating with sample DNA concentration due to Illumina new patterned flowcell. Additionally, we observed other characteristics that were typical of sequencing platfor We observed how error rate and quality varied by cycle to learn typical patterns of failed samples. Finally, we found two distinct patterns of coverage over %GC for each library type: TruSeq PCR free, which retains high coverage in extreme %GC areas (≥80%), as compared to TruSeq Nano v2, which has steady coverage in 30-70% GC.

Although these metrics are informative, there may be more useful elements to assess sample performance in a production pipeline. Standard examples include whole genome and coding region coverage statistics. In addition, we found quantifying the base pair coverage fluctuations to be crucial in determining coverage uniformity across large regions of the genome. Finally, we have done variant comparisons to assess the sensitivity and specificity of the platform, as well as variant reproducibility in technical replicates. Scrutinized together, coverage and variant analysis should provide insight to the tipping point in which depth and uniformity of coverage start to affect the ability to detect true variants.

An intra-host longitudinal genomic comparison of Burkholderia pseudomallei provides insight into long term adaptation

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.16

Crystal Hepp[1], Jason Sahl[2], Heidie Hornstra[1], Karthik Handady[1], Christopher Allender[1], Erik Settles[1], Apichai Tuanyok[3], Erin Price[4], Mirjam Kaestli[4], Mark Mayo[4], Richard Bowen[5], Timothy Kidd[6], Scott Bell[7], Paul Keim[1], Talima Pearson[1], Bart Currie[4]
[1]Northern Arizona University, [2]Translational Genomics Research Institute, [3]University of Hawaii, Manoa, [4]Menzies School of Health Research, [5]Colorado State University, [6]The University of Queensland, [7]The Prince Charles Hospital

There are numerous examples of soil microorganisms that have made the transition from saprophyte to human pathogen, including Bacillus anthracis, Coxiella burnetii and Burkholderia mallei. With this transition between ecological niches comes a sharp contrast in selective pressures, including the availability of nutrient resources and diverse antimicrobial opposition. Therefore, the expectation is that extensive host-specific adaptations would take place in the pathogen population. This hypothesis should hold true particularly in the case of a chronic infection, which provides the unique opportunity to witness the evolutionary mechanisms allowing for a pathogen to become well-suited to its human host.

Here, we present a longitudinal investigation of a B. pseudomallei infection that initially presented as severe respiratory melioidosis. Interestingly, after antibiotic treatment, lung lobectomy, and 13 years, the patient has become the first ever asymptomatic chronic carrier of B. pseudomallei. In order to determine the within-host bacterial adaptations that have occurred, we performed molecular evolutionary analyses comparing both morphologies and whole genomes of 100 intra-patient and linked environmental isolates to each other as well as to other closely related strains and species. Longitudinal studies on bacteria rarely incorporate this amount of information, but allow for a more complete picture of the diversification within and among time points and biological compartments.

Intra-host evolutionary dynamics have given rise to fluctuations in this bacterial quasispecies that contribute to antibiotic resistance, changes in growth rate, multiple deletions up to 150,000 bp in length, and a reduction in virulence allowing for long-term establishment. Mapping these events back to a phylogeny has revealed that while many events occurred only once, some adaptations have arisen independently multiple times. Additionally, comparison of this population to B. mallei, which is an equine-adapted clone of B. pseudomallei, exposes numerous instances of parallel evolution that have occurred on the path to becoming host-adapted. Finally, a Bayesian analysis of population dynamics over time shows an increase of the effective population size that coincides with the patient’s most severe duration of symptoms.

From FinisherSC to MetaFinisherSC: tools to upgrade de-novo assembly using long reads

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.17

Ka-Kit Lam[1]
[1]***University of California Berkeley***

We introduce FinisherSC, which is a repeat-aware and scalable tool for upgrading de-novo assembly using long reads. Experiments with real data suggest that FinisherSC can provide longer and higher quality contigs than existing tools while maintaining high concordance. Moreover, we describe our initial work on extending FinisherSC to MetaFinisherSC which targets improving metagenomics assemblies.

Next Generation Sequencing Sample Preparation Utilizing the Echo® Liquid Handler

Wednesday, 27th May 20.00 - Santa Fe Room (Sponsored by Roche) – Poster 1b.18

***Howard Lee*[1], *Danny Lee*[2], *Kelvin Chan*[2]**
[1]*LabCyte*, [2]*SeqMatic*

The advent of Next-generation sequencing (NGS) has enabled researchers to overcome the limitations in resolution, scalability, and throughput experienced with capillary electrophoresis-based Sanger sequencing. While these technological advances have lowered the cost of sequencing, upstream library preparation remains a significant bottleneck and a prime target for automated liquid handling. The ability of Echo liquid handlers to acoustically transfer samples and reagents without tips or contact provides an efficient, contamination-free solution for genomic library preparation. The precision and accuracy of sub-microliter transfers from any microplate well to any microplate well accelerates and improves library pooling and normalization with less setup time in comparison to methods utilizing manual pipetting. In this work, the Echo 555 liquid handler was used to prepare libraries produced from E.coli for sequencing with the Illumina® MiSeq sequencer.

STR-Seq: highly parallel short tandem repeat analysis method based on programmable target selection

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 1b.19

***Giwon Shin*[1], *Hojoon Lee*[1], *Sue Grimes*[1], *Erik Hopmans*[1], *Hanlee Ji*[1]**
[1]*Stanford University*

Short tandem repeats (STR) are a type of genetic variation which has the fastest mutation rate. As a result, it is one of the most informative parts of genome, of which the most representative application is forensics DNA fingerprinting. More interestingly, its clinical implications have recently been highlighted particularly in studies dealing with genetic evolution of cancer cells. Despite their wide application in a variety of fields, however, the analysis of STRs faces two major technological issues when analyzed with next generation sequencing-based methods: Only the reads which encompass an entire STR locus are informative, and PCR amplification during library preparation can introduce stutter noise. Consequently, analysis on STRs requires more sequencing depth with finely controlled target selection, but current target enrichment methods are not appropriate for this purpose. In this study, we developed a novel targeted sequencing technology, short tandem repeat sequencing (STR-Seq), by which sequencing resources can be used to generate only the STR-spanning reads. To improve the target selection specificity, we programmed Illumina sequencing flowcell to capture only the informative libraries. Unlike conventional methods, capture probes are immobilized and physically separated on the flowcell, which maximize the efficiency and specificity. As a result, we could demonstrate simultaneous assay of approximately 2,500 STR loci. In addition, to eliminate stutter noise, sequencing libraries were prepared by an amplification-free method which has extremely high adapter ligation efficiency. In the data analysis process, we developed a method of counting STR repeat motifs and associated variants both internal and external of the STR. Furthermore, we demonstrate that this technology has concordance with conventional methods while the measurement is generated in much higher throughput.

Lower Cost, Higher Throughput Library Preparation with the Echo liquid handler® and the NuGEN Ovation® Single Cell RNA-Seq System

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 1b.20

John Lesnick[1], Jon Don Heath[2], I Ching Wang[2], Marie Eide[2], Steven Kain[2]
[1]LabCyte, [2]SeqMatic

The rapid evolution of next-generation sequencing (NGS) technologies is accelerating our knowledge of gene expression, regulation and pathway complexities in mammalian cells. Transcriptome analysis with NGS offers increased transcript coverage to enable the detection of rare transcripts, novel alternative splice isoforms and the measurement of transcript abundance. However, traditional library preparation methods for NGS are often not amenable to transcriptome analysis. Traditional methods carry a requirement for a large amount of total RNA to yield sufficient mRNA to analyze which can be unobtainable or cost prohibitive in most experiments. The NuGEN Ovation Single Cell RNA-seq System is a highly sensitive and complete library preparation procedure for whole-transcriptome sequencing that requires total RNA from samples as small as a single cell or 10 picograIn this proof of concept study (POC), the Ovation Single Cell RNA-Seq System was validated by examining libraries prepared and sequenced at full and miniaturized reaction volumes. Furthermore, the sample and reagent transfers were automated using the Echo acoustic liquid handling technology. Echo liquid handlers transfer a wide range of fluids without contact of tips or recalibration between fluid types. The industry leading accuracy and precision of Echo liquid handlers at microliter and nanoliter volumes in combination with the NuGEN Ovation Single Cell RNA-Seq System increases library preparation throughput while reducing the costs to enable a broader application of transcriptome analysis with NGS.

Read depth and seedling number influence on the development of De Novo transcriptomes for Leucaena tree species

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.01

Diana Dugas[1], Madhugiri Nageswara-Rao[1], Richard Cronn[2], Donovan Bailey[1]
[1]New Mexico State University, [2]USDA Forest Service, Pacific Northwest Research Station

As next generation sequencing becomes ever cheaper, efficient, and available, more research is able to focus on non-model organisThe relatively low cost and focus on functional elements of the genome have resulted in transcriptome studies becoming popular for research on non-model organisHowever, the optimization of sample selection and read-depth has not been standardized, yet they have a profound impact on the resulting transcriptomes. Here we present data on two different out-crossing, non-model species, Leucaena trichandra and L. cruziana, belonging to the Leguminosae family. We generated Illumina libraries and HiSeq 100bp paired-end reads for a range of seedling numbers, which were used to assemble de novo transcriptomes. The results suggest that transcriptome assembly from a single seedling is not as comprehensive as those for higher order seedling number from the same maternal line. However, no significant difference is found among transcriptomes based on 3, 5, and 8 seedlings. We also explore the number of reads necessary to generate a comprehensive transcriptome assembly by subsampling the available reads. We found that 37M read-pairs are sufficient to reconstruct the more complex transcriptome, L. cruziana. Restricting the read-pairs to 31M or 24M decreases the number of transcripts recovered and projected gene number, as well as gene completeness as demonstrated by full-length orthologs found in the TRAPID database. In conclusion, we offer some guidance for other researchers who wish to assemble de novo transcriptomes in non-model plants, both by presenting our method for determining best parameters and by suggesting a starting experiment.

Comparisons between Illumina and Pacific Biosciences microbial genome assemblies and evaluation of DNA regions intractable to next generation sequencing

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.02

***Sagar Utturkar**[1], **Richard Hurt**[2], **Dawn Klingeman**[2], **Steven Brown**[2]*

[1]University of Tennessee, [2]Oak Ridge National Laboratory

Background:

Development of next generation sequencing (NGS) technologies has revolutionized genomics research by providing high-throughput, low-cost sequencing methods. Despite sequencing and assembly advances, there are several examples of microbial genomes which remained unfinished by PacBio and Illumina platforThe aim of the present study was to reveal and characterize regions of DNA intractable or unresolved by either Illumina or PacBio technologies.

Methods:

The genomes of Clostridium thermocellum AD2, Clostridium pasteurianum ATCC 6013, Clostridium autoethanogenum DSM 10061, Clostridium paradoxum JW/YL-7T, Pelosinus fermentans UFO1, Pelosinus fermentans JBW45, Halomonas sp. KO116 and Bacteroides cellulosolvens DSM 2933 were sequenced using combination of Illumina paired-end (PE) and Pacbio RS-II platforDe novo and hybrid assemblies were performed with only Illumina, only PacBio and Illumina + PacBio data combinations using SPAdes, ABySS and SMRTanalysis software. Complete genome assemblies generated by PacBio data were compared with Illumina draft assemblies to reveal genomic regions which were intractable by Illumina technology. From the genomes which could not be resolved completely by PacBio and Illumina technology, the genome of B. cellulosolvens was finished by following PCR and Sanger sequencing approach. The manually finished B. cellulosolvens genome was compared to the previous best assembly to examine DNA regions which were not determined by NGS. In silico characterization of these intractable regions was performed to identify the properties such as ability to form DNA hairpin-loop structures, associated free energy (ΔG) and annotations.

Results:

Our results indicated that DNA regions intractable by Illumina technology were mostly associated with large repeats in the genome such as rRNA operons. The regions which were intractable by PacBio technology were mostly associated with repetitive regions as well as characterized by the potential to generate strong DNA hairpin-loop structures. In this study, PacBio only assemblies had better statistics compared to Illumina only or hybrid assemblies. Gene number comparison of draft and finished genome revealed several new and/or longer open reading frames in improved genomes.

Conclusion:

In our study, PacBio technology generated assemblies with the best summary statistics. DNA regions intractable to Illumina and PacBio technology were largely associated with rRNA operons and DNA hairpin-loop structures, respectively.

Acknowledgement:

Research sponsored in part by the Genomic Science Program, U.S. Department of Energy, Office of Science, Biological and Environmental Research and by LanzaTech NZ Ltd.

Changes To NGS Workflow…Why Accurate Sizing And Quantification Of Library Preparations Are Critical To Successful Sequencing

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.03

Steve Siembieda[1]***, Jon Hagopian***[1]***, Jolita Uthe***[1]***, Kit-Sum Wong***[1]
[1]***Advanced Analytical***

Labs doing NGS library prep are confronted with many challenges each day including varying incoming sample quality, sample prep loss, pooling/barcoding issues, quantification inconsistencies and fluctuating sample numbers to name a few. Propitiously the major sequencing platforms have enhanced their capabilities allowing the use of suboptimal, small sized and/or low quantity DNA and RNA raw materials. This in turn has opened up new and previously unimagined ways to generate sequencing data. With these advances, however, inaccurate quantification of library preps made from poor quality raw materials, from pooled libraries made from different samples or unbalanced barcoding strategies can affect both sequencing results and radically affect costs. Back in the early days of NGS, few NGS library prep methods were available. Today there are many ways to produce libraries. A better understanding of the raw and finished nucleic acid materials are needed because labs face both constraining challenges imposed on them by working with low quality nucleic acids and enabling challenges that open the possibility of doing longer fragment length reads. Due to these factors, instrumentation which can sensitively and accurately assess the quality and quantity of nucleic acids can improve the consistency of results generation. The Fragment Analyzer™ from Advanced Analytical was purposely designed to meet the challenges of NGS library prep assessment. By utilizing the superior properties of capillary electrophoresis, the Fragment Analyzer™ can effectively separate fragment sizes from very small (microRNA) to very large (genomic DNA) and all sizes in between as well as provide high quantification accuracy. Information will be presented on the advances made to library preparations methods and why accurate concentration and size assessment is important to achieving good sequencing data and final results derived from both short and long read sequencing platforms.

BioVelocity is a comprehensive and rapid post-sequencing bioinformatics tool to enhance food biosecurity, biosurveillance and outbreak investigations through precise pathogen detection

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.04

Jon Kennedy[1], Danielle Montoya[1], Jane Tang[1], Karen Taylor[1]
[1]Noblis

Food biosecurity is an element of biodefense, which includes protecting our food supply from bioterrorism as well as from accidental or natural contamination. As the cost of whole genome sequencing (WGS) falls, it is increasingly feasible to use it for real-time outbreak prevention and biosurveillance. Traditional microbiological detection methods require days or even weeks to identify the causative agent. BioVelocity, a Noblis-developed post-sequencing bioinformatics tool, utilizes single nucleotide polymorphisms (SNPs) from whole genome sequences to rapidly identify unknown pathogens at the strain level. Currently, BioVelocity performs 100x faster than the industry standard for sequence read alignment. With BioVelocity, sequenced reads are directly subjected to analysis without requiring assembly or annotation, which greatly increases speed while maintaining fidelity. BioVelocity can align raw sequence reads, perform SNP detection and pathogen identification, and even metagenomic analysis in 23 seconds (per genome) or less for average-size microbiological pathogens. BioVelocity's unique indexing technology allows for unprecedented speed in identification of genomic regions that can be used as references for geographical association as long as the metadata are available. Additionally, relationships between serovars that frequently cause outbreaks every year can be determined using WGS.

In 2012, a nationwide outbreak of Salmonella bredeney occurred stemming from Valencia peanut butter products. We previously demonstrated the power of our approach by applying BioVelocity to thousands of Salmonella sequences present in the Food and Drug Administration's BioProject. From there through a phylogenetic approach we were able to identify the cluster representing the 2012 outbreak and, for the first time, inferred additional food isolates in the database that were likely to have been related to the outbreak and, as far as we could determine were not previously associated as such. In this analysis, BioVelocity was used to analyze 103 genomes directly related to peanut butter food sources and specifically the 2012 outbreak in 40 minutes (23.3 seconds/genome). Noblis' work with this dataset demonstrates how BioVelocity can rapidly and accurately cluster, trace, and make inferences based on our results and accompanied sample metadata.

Noblis is in the early stages of a pilot that would allow food producers or others in the food industry to submit data from their food safety testing programs to be sequenced and analyzed. Our goal is to help the food industry see the value of WGS to reduce the risks and costs of contaminated foods, improve food safety plans and procedures, and inform capital investment decisions. Additionally, our results indicate that BioVelocity has potential for improving food biosecurity for all by accurately detecting pathogens, quickly tracing outbreaks, and even differentiating between intentional or natural outbreaks through biosurveillance using our unique and efficient approach.

Smart RNA Sequencing (SRS): An Efficient NGS Method For Improved Diagnostic and Transcriptomic Applications

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.05

Cheryl Gleasner[1], Kim McMurry[1], Andrew Hatch[1], Cheriece Margiotta[1], James Horne[1], Patrick Chain[1], Jason Gans[1], Momchilo Vuyisich[1]

[1]Los Alamos National Laboratory

We want to transform the field of infectious disease diagnostics with Smart RNA Sequencing (SRS) biotechnology platform that enables efficient use of next generation sequencing (NGS). SRS improves the efficiency of NGS by enriching pathogen RNAs in clinical samples. This is achieved by specifically removing a large fraction of the most abundant, but non-informative, background RNAs. All other RNAs (including those from pathogens) are enriched. For example, removing 90% of the non-informative RNA provides a 10-fold enrichment of pathogen RNA. In turn, this enables 10 times less expensive (and faster) sequencing.

SRS also includes robust, inexpensive and rapid sample lysis, RNA extraction, and library preparation steps. It has been shown that in multiplexed sequencing runs, false positive results can occur due to sample-to-sample cross-talk. The dual barcode method that we have recently employed in SRS has the potential to virtually eliminate false positive sequencing results by incorporating an additional parameter to verify results. In addition to the laboratory steps, a custom data analysis pipeline developed at LANL is being used to rapidly detect all pathogens.

Effects of Wood Smoke Exposure on the Oropharyngeal Microbiome

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.06

Arthur Armijo[1], Kimberly Paffett[2], Kevin Harrod[3], Darrell Dinwiddie[2]
[1]University of New Mexico, [2]University of New Mexico Health Sciences Center, [3]University of Alabama Birmingham School of Medicine

Objective/Specific Aims: Recent human microbiome studies have revealed the importance of the microbiome as a biomarker for human health and disease. Specifically, alterations in the microbiome have been found to be associated with both increased risk of infections and in response to infections. The objective of this study is to elucidate the impact of wood smoke exposure on the human oropharyngeal microbiome and to examine if wood smoke exposure alters colonization density of potentially pathogenic organis

Methods/ Study Population: We recruited 16 individuals with asymptomatic colonization of Streptococcus pneumoniae for randomization to exposure to 4 days of filtered air or hardwood smoke generated from a contemporary wood stove. Oropharyngeal swabs were collected prior to exposure, after each of 4 consecutive days of exposure, and 1 week and 28 days following the initial exposure. Isolated DNA was enriched for variable regions 4-6 of the 16s rRNA gene by PCR and sequenced 2x300bp on the Illumina MiSeq. Data analysis was performed using the Illumina BaseSpace 16S Metagenomics v1.0 and the Kraken Metagenomics bioinformatic pipelines.

Results: Wood smoke exposure resulted in shifts in both the diversity of the microbiome and abundance of specific genera. After 4 days of exposure, 9 of 10 subjects exposed to wood smoke had significant increases in percent abundance of at least one genus of potentially pathogenic bacteria. Streptococcus was significantly increased in 6 of 10 subjects, Neisseria was increased in 5 of 10 subjects and one subject had a significant increase in Actinobacillus. Only 1 of 6 subjects exposed to air had a significant increase identified (Veillonella).

Discussion/Significance of Impact: Our results suggest that wood smoke exposure significantly alters the oropharyngeal microbiome and specifically increases the abundance of potentially pathogenic organisAdditional studies are needed to further examine the impact of exposure to environmental pollutants, such as wood smoke, on the human microbiome and how exposures may increase the risk for infection, particularly from colonized organisms.

Whole Genome Sequencing with HiSeq X Ten System at The Genome Institute, Washington University School of Medicine

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.07

Aye Wollam[1], Lisa Cook[1], Bob Fulton[1], Lucinda Fulton[1], Richard Wilson[1]
[1]The Genome Institute at Washington University in St. Louis

When used at full-scale, the Illumina HiSeq X Ten system enables whole genome sequencing at an unparalleled scale by generating massive data quantities in a quick turn-around time while aiming to break the $1000 per genome cost barrier. At The Genome Institute, our process optimization focuses on obtaining high yield/low error rate sequence while being cognizant of factors influencing those metrics. Library construction methods such as Illumina TruSeq Nano DNA kit, Illumina TruSeq PCR-Free DNA Kit, and other vendor kits, as well as various pooling methods and load concentrations are being evaluated as to their impact on the quality, quantity and representation metrics. Here we present some of these results and identify key factors associated with the performance of these new instruments.

Evaluation of Life Technology's Human Identity SNP Kit for the Ion Torrent™ PGM™ Sequencer

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.08

James Robertson[1], Kelly Meiklejohn[1]
[1]FBI Academy / FBI Laboratory

In cases where only a partial or incomplete STR profile is obtained from a sample, information contained in single nucleotide polymorphisms (SNPs) can prove informative for human identification. Life Technologies recently released two SNP multiplex panels compatible with their high throughput Ion Torrent™ PGM™ sequencer: the HID-Ion Ampliseq™ Identity Panel and the HID-Ion Ampliseq™ Ancestry Panel. In this study we evaluated the reproducibility and sensitivity of the Identity Panel, which contains primers for the amplification of 90 autosomal SNPs and 34 Y-clade SNPs. This was completed in three phases, without deviating from the manufacturers' published protocol: (1) for five commercially available pure native DNAs, libraries were individually prepared six separate times using 1 ng of DNA (n=30), (2) for two of the pure native DNAs, six additional libraries were prepared using both 0.2 ng and 0.5 ng of DNA (n=24), and (3) using three forensic type samples (saliva, blood and body fluid) from both a male and female donor, libraries were prepared in triplicate using both 0.5 ng and 1.0 ng of DNA (n= 36). We will discuss the library yields, percent of templated ISPs and general sequencing metrics obtained, and also report on the coverage/quality and congruence for all SNPs in the panel.

Whole Genome Sequencing of Recurrent Methicillin-Resistant Staphylococcus aureus Enables High Resolution Genotyping

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.09

Murtada Alsaadi*[1], Kimberly Paffett[1], Walter Dehority[1], Jon Femling[1], Renee-Claude Mercier[1], Darrell Dinwiddie[1]
[1]University of New Mexico Health Sciences Center

Over six million people are naturally colonized with methicillin-resistant Staphylococcus aureus (MRSA) and an estimated 75,000 active infections occur each year in the United States causing significant morbidity, mortality, and financial burden on the healthcare system. Recurrent infections account for up to forty percent of MRSA cases. Determining the genetic composition and evolution of recurrent MRSA may provide critical information to aid in its clinical management and prevention. To this end, we undertook shotgun whole genome sequencing (WGS) of 16 samples from 6 patients with recurrent MRSA infections of the blood. Sequencing of 2x75bp on the Illumina MiSeq resulted in mean whole genome coverage of 50-80X. Sequences were aligned to three prototype complete genome reference strains NC_007793 (USA 300), NC_002745 (USA 100), and NC_007795 using the CLC Bio Genomics Workbench. Whole genome sequencing-based strain typing revealed a 25% discordance rate with previous pulse-filled gel electrophoresis (PFGE) typing. Subsequent variant analysis of recurrent cases uncovered genetic variation in strains isolated from the secondary infection, which may represent evolution of clonal MRSA. Our results indicate that WGS provides higher resolution in identification of MRSA strains as compared to current standard of care PFGE. Furthermore, WGS enables single nucleotide variant detection that can provide insight into the evolution, adaptation, and acquisition of mutations in MRSA during the course of clinical infection. Together, this genomic data has the potential to improve clinical management of pathogenic infections.

Assessing Three Whole Metagenome Analysis Tools for Their Ability to Identify Burkholderia pseudomallei Within the Pulmonary Microbial Community of Melioidosis Patients

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.10

***John Gillece**[1], **Jim Schupp**[2], **Jason Sahl**[2], **Rebecca Colman**[2], **Jordan Buchhagen**[2], **Hannah Heaton**[2], **Josie Delisle**[2], **Bart Currie**[3], **Mark Mayo**[3], **Paul Keim**[4], **Dave Wagner**[4], **David Engelthaler**[2]*

[1]TGen North, [2]Translational Genomics Research Institute, [3]Menzies School of Health Research, [4]Northern Arizona University

In light of recent advances in high throughput sequencing, Whole Metagenome Sequencing and Analysis (WMSA) has received attention because of its potential as a tool for rapid diagnosis of infectious agents from clinical samples. It would obviate the need for upfront diagnostic assays or culture techniques that require prior knowledge and could detect a broad spectrum of targets from a single metagenomic library, including viral, fungal, and bacterial. Nonetheless, while generating the data necessary to identify the targets is achievable, parsing these data with sensitivity, specificity, and accuracy is not trivial. In this study, WMSA was applied to six sputum samples and one pleural fluid sample from active melioidosis patients and three tools designed for such datasets were assessed: Surpi, Metaphlan, and Lasergene. Melioidosis, caused by Burkholderia pseudomallei, is a public health and potential bioterrorism agent endemic to Southeast Asia and Northern Australia. The metagenomes were sequenced on the Illumina HiSeq, resulting in 30M-97M paired reads. Surpi and Lasergene operate similarly, aligning reads against large databases after filtering human and ribosomal RNA matches with the ability to detect fungal, bacterial, and viral. Metaphlan differs in that it utilizes clade specific markers and, at this time, only for bacterial organis From less than 100 reads to tens of thousands of reads mapped in a dispersed fashion to B. pseudomallei from each of the samples, with the pleural fluid sample having over 15k map using Lasergene. For the same sample, Metaphlan, which reports in relative abundance, identified Burkholderia pseudomallei as comprising 95% of the entire microbial community detected in the sample. Metagenomic analysis also indicated that the pleural sample had <25 reads map to any other bacterial genome. The sputum samples had much more diverse communities, with only one having B. pseudomallei as the dominant organism. The remaining samples had common commensals such as Prevotella, Veillonella, and Rothia at varying levels of abundance. In 3/6 sputum samples B. pseudomallei was in the top 10 most abundant organisThese samples tended to harbor other possible pathogenic agents in the top 10 most abundant organisms, such as Pseudomonas aeruginosa, Streptococcus pneumoniae, and Klebsiella pneumoniae, while samples with relative low B. pseudomalleil abundance tended to harbor primarily commensals in the top 10 organis In the end, Surpi and Lasergene performed comparably in both their high level of sensitivity but low level of specificity. On the other hand, Metaphlan was at least twice as fast as the other tools and quite specific; however, because it relies on clade specific markers rather than whole genome sequences, it likely was unable to detect targets that are present in the sample at low levels.

Improving Pathogen Detection and Characterization through Next Generation Sequencing and Metagenomics

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.11

Chris Hopkins[1], Eishita Tyagi[1], Scott Burns[1], Cliff McDonald[2], Duncan MacCannell[2], Jamie Posey[2]

[1]Centers for Disease Control and Prevention, Atlanta, GA / Booz Allen Hamilton, [2]Centers for Disease Control and Prevention, Atlanta, GA

Infectious disease detection, characterization, and surveillance for public health traditionally rely on methodologies that identify etiologic agents based on previously characterized associations to known clinical syndromes. In the case of novel emerging pathogens, however, attempting to detect previously uncharacterized infectious agents can prove extremely challenging. The field of metagenomics has the potential to revolutionize pathogen detection in public health laboratories through the use of next generation sequencing (NGS) on primary clinical samples. Metagenomics allows for the simultaneous detection of all microorganisms in a sample without a priori knowledge of their identities. However, a significant challenge with this approach is the vast majority of NGS information generated from a single complex clinical specimen is derived from the host genome and/or commensal organisms (i.e.,"the noise"); only a small remainder of the sequencing information can be truly utilized for identification and characterization of the underlying disease-causing pathogen (i.e.,"the signal"). Thus, the primary challenge is to increase this signal-to-noise ratio. Despite sequencing of isolated organisms becoming a routine task, the laboratory and bioinformatics methods required to characterize pathogens within complex, heterogeneous clinical matrices (e.g., blood, sputum, stool, cerebrospinal fluid, tissue, etc.) are still in development. Successful clutter mitigation – the separation of genetic material of potential pathogens from the genomes of background species or host organisms – may improve pathogen identification and characterization of microbial communities from complex samples. In the context of public health surveillance, where the increasing clinical adoption of culture independent diagnostics threatens the availability of isolated pathogens, clutter mitigation will be a necessary first step when processing nucleic acid from specimens. The effectiveness of clutter mitigation techniques in clinically relevant specimens has not yet been compared in a systematic fashion. Using previously characterized specimen sets (blood, sputum, and stool), we are assessing a myriad of CDC-developed and commercially available host/commensal depletion and targeted enrichment strategies that chemically, physically, or enzymatically increase the signal-to-noise ratio. In this study, we used NGS and qPCR to assess the impact of DNA and RNA based clutter mitigation technologies. DNA-based host/commensal depletion strategies resulted in up to 4-fold enrichment of non-human genomic material in blood specimens and up to13-fold enrichment in sputum specimens. Using host ribosomal RNA depletion strategies, we observed up to 6-fold enrichment of non-host ribosomal RNA in blood specimens. Additionally, using targeted enrichment strategies we observed near complete pathogen genome coverage and enrichment above background at levels less than 1% of known spiked pathogen DNA in sputum specimens. The resulting clutter mitigation protocols from this systematic study will ultimately be applied to real-world outbreak samples to assist public health programs to produce and analyze clinical data for improved pathogen detection and characterization.

Using high density SNP data for QTL mapping and making improvements to genomic scaffolds

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.12

Sheina Sim[1], Bernarda Calla[2], Scott Geib[2]
[1]University of Hawaii, Manoa, [2]USDA-ARS Daniel K. Inouye US PBARC

The Mediterranean fruit fly Ceratitis capitata (Wiedemann), commonly known as the medfly, is a destructive agricultural pest and the object of expensive population eradication and suppression efforts within state and federal departments of agriculture. Area-wide integrated pest management programs control medfly populations through the release of sterile males which must be massively produced. Mass-rearing and release of sterile males is facilitated by two sex-linked traits white pupae (wp) and temperature sensitive lethal (tsl). Though these two sex-linked traits in what is known as a genetic sexing strain were developed over 20 years ago, their genetic basis is unknown. To map one of these traits, wp, three mapping populations were generated from an F4 wp strain and wild type lab line strain cross. Three sibships with individuals segregating for the white pupae trait were then genotyped by sequencing (GBS) and identifying SNPs from a restriction site associated DNA library. The SNP genotypes were then used to identify chromosome-scale linkage groups and linked loci were identified using quantitative trait loci (QTL) analysis. This data has been used to develop a genetic assay for differentiating between recaptured sterile insect released males and wild males, will be useful for the improvement of existing sterile insect colonies, and can be used to identify an orthologous gene in other species to create novel sterile insect strains.

High Quality Library Construction and Reliable Quantitation with NEBNext Reagents

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.13

Pingfang Liu[1], Nathan Tanner[1], Janine Borgaro[1], Lynne Apone[1], Vaishnavi Panchapakesa[1], Deyra Rodriguez[1], Erbay Yigit[1], Brad Langhorst[1], Don Johnson[1], Julie Menin[1], Christine Sumner[1], Christine Chater[1], Fiona Stewart[1], Nicole Nichols[1], Eileen Dimalanta[1], Theodore Davis[1]

[1]New England Biolabs, Inc.

The field of NGS has matured significantly over the past few years. While the use of NGS in the clinic was simply a possibility a few years ago, today it is routinely used in some areas of diagnostics. An expanded role for NGS in the clinic and research will depend on continued improvement of the upstream processes required to produce high quality NGS data. In particular, maximizing data output & minimizing instrument run failure are imperative. In this poster, we present data showing: 1) the improvements we've made in library preparation with the development of the NEBNext Ultra II Library Prep kit; and 2) the development of the NEBNext Library Quant Kit for Illumina, a simple and robust method for quantitation of Illumina libraries. In the first part, we show that libraries made with the Ultra II kit have low DNA input requirements; significantly improved library yields and reduced sequence bias. In addition, the workflow is simple and streamlined, greatly reducing the time required to produce high quality DNA libraries and the possibility of errors. In the second part, we demonstrate the effectiveness of the NEBNext Library Quant Kit for a broad range of library types and sizes as well as advantages offered by qPCR quantitation for obtaining optimal cluster density and user-to-user consistency. The NEBNext Quant Kit offers an efficient and cost-effective qPCR library quantitation workflow for users looking to optimize both sequencing yield and throughput.

Unsupervised Phylogeny with Automatic Correction for Horizontal Gene Transfer and Optional Filtering of Mobile Elements

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.14

Raquel Bromberg[1], Zbyszek Otwinowski[1]
[1]**University of Texas Southwestern Medical Center**

Advances in sequencing have generated a large number of full genomes. In the past, phylogenetic analysis was mainly performed with alignment-based methods, which are generally not scalable and require a set of orthologs. The definition of these orthologs is nontrivial, and the process of separating them from the paralogues that are generated by horizontal gene transfer and gene duplication followed by selective loss is not necessarily well-defined, as in the case of Aquifex aeolicus. Methods that take whole genomes or proteomes as their input are now coming into fashion, in particular due to their ease of use and scalability, but often a large fraction of the genes of a prokaryote are mobile elements that do not reflect evolution by descent.

We have developed SlopeTree, a method which produces phylogenies constructed from whole proteomes and is therefore free from dependence on orthologue identification. SlopeTree both identifies and corrects for horizontal gene transfer at multiple stages, making it more robust than past efforts in this area. Using the statistics of exact kmer matches between proteomes, the method automatically and efficiently calculates genomic evolutionary distances, with robust phylogenic inference even when using a fast, Neighbour-Joining method. Included in this calculation is a multi-level correction for the effects of horizontal gene transfer, as well as corrections for composition and low complexity sequences, and nonlinearity of mutation accumulations. In addition, SlopeTree provides an optional feature for filtering out mobile elements prior to the main run. This filter is adjustable and can automatically separate core proteins from mobile elements.

For 495 bacteria, 72 archaea, and 72 strains of Escherichia coli/Shigella, a range of trees was generated using as input both whole proteomes and proteomes with mobile elements filtered out at different levels. These trees were compared with the NCBI taxonomy and trees produced by conserved protein concatenation. In general, SlopeTree generates sensible topologies which are relatively stable between whole proteome and reduced proteome inputs, which validates the concept of species and phyla as having a core proteome evolving by descent, but not necessarily coevolving with the ribosome and its proteins.

The UVRI genome center: a model for successful collaborative capacity building at the Uganda Virus Research Institute

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.15

John Kayiwa*[1], *Jonathan Kayondo*[1], Timothy Bayaruhanga[1], Julius Lutwama[1], Jeff Borchert[2], Trevor Shoemaker[2], Tracy Erkkila[3], Momochilo Vuyisich[3]
[1]Uganda Virus Research Institute, [2]CDC/Uganda Virus Research Institute, [3]Los Alamos National Laboratory

The Uganda Virus Research Institute (UVRI) is a national reference laboratory for a number of various viral diseases including HIV, Influenza and Hemorrhagic fevers. Next generation sequencing (NGS) is slated for incorporation, at UVRI, into laboratory diagnosis especially during epidemic outbreak investigations for enhanced pathogen detection. To this end, a regional genome center with capacity for deep sequencing , and for use by the Government of the Republic of Uganda under the Cooperative Biological Engagement Program (CBEP), is being established at the UVRI National Influenza Center (NIC) with assistance from various collaborating and development partners including the Los Alamos National laboratory, The Defence Threat Reduction Agency (DTRA) operating under the Department of Defence (DoD) of the United States of America, and Centers for Disease Prevention and Control (CDC). An Illumina MiSeq sequencer with accompanying computing infrastructure to support bioinformatics analysis has been installed at the institute. In addition, complementary training to 3 core staff in sample processing, sequencing and data analysis is being provided. We hereby share methods and data from recent meta-transcriptomic analysis test runs at the facility.

Four RNA samples, ARBO_0468UVRI, ARBO_0473UVRI, ARBO_0474UVRI and ARBO_0478UVRI isolated from whole blood of patients presenting with various viral hemorrhagic fever (VHF) symptoms were processed for transcriptome sequencing. The initial RNA was quality control checked using Qubit Quantification Protocol (RNA assay) and Bioanalyzer RNA Chip from Agilent Technologies. The samples were depleted of Human rRNA using Ribo-Zero Magnetic Kit rRNA removal protocol and the remaining mRNA concentration was measured using Qubit. Library preparations were ligated to indicies and adapter sequences of the fragmented RNA following the Epicentre's ScriptSeq v2 sample Preparation protocol. Library validation was by dsDNA HS Assay Qubit Quantification Protocol. Accurate quantification of amplifiable libraries was calculated to obtain the final library concentration required for cluster generation and sequencing on the Ilumina Miseq. Next –generation sequencing was performed according to Illumina Miseq Protocol.

Up to 13 million reads per sample were generated and imported into both CLC bio Ver 7.5.1 and EDGE for quality check, cleaning and community profiling. Cleaning/trimming included, removal of host and adapter sequences; removal of poor quality sequences (>0.01 Q-score limit); and discarding short reads (<50bases). CLC and EDGE metagenomic analysis outputs of patient sequences lacked substantial blood-stage gene expression, even at 0.01% detection level, from known agents causing VHF or other severe clinical disease, perhaps indicating a need for supplementary diagnosis from other sample types. Other than ARBO_0478 that showed Haemophilus parainfluenzae T3T1 at 0.02% of all expressed sequences, the rest had either non-pathogenic or opportunistic microbes such as (Pseudomonas putida, Chlorobium phaeobacteroides, Pandoravirus salinus) repeatedly appearing among the top 10 most highly expressed. More in-depth bioinformatics manipulation such as exclusion of non-unique sequences, such as tRNA loci, from the analysis could improve accuracy of our metagenomic outputs.

An Intuitive '16S rRNA Biodiversity Tool'

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.16

Christian Olsen[1], Chris Duran[1], Richard Moir[1], Kashef Qaadri[1], Helen Shearman[1], Alex Cooper[1], Brett Ammundsen[1]

[1]Biomatters

The Geneious R8 ‘16S rRNA Biodiversity Tool’ is a cloud-based tool for routine species classification and relative abundance measurement using high throughput 16S rRNA amplicon sequencing data from environmental samples. Preprocessed full-length bacterial 16S rRNA sequences may be utilized, but any sub-region of 16S can be used.

The user submits their next-generation sequence data through the Geneious R8 bioinformatics platform to a distributed cloud compute resource. The data are then analyzed using the Ribosomal Database Project (RDP) database Classifier. The RDP Classifier assigns sequences derived from bacterial and archaeal 16S genes and fungal 28S gene to the corresponding taxonomy model using a ‘Naïve Bayesian Classifier’ for rapid assignment of rRNA sequences.

The Geneious ‘16S Biodiversity Tool’ accurately assigns a taxonomy (in the range of domain to genus) along with a confidence-estimate for each sequence by comparing them to the RDP database. A secure weblink is returned within Geneious’ ‘Document Viewer’. Upon clicking the weblink, the output is then displayed in a web browser using Krona, which produces an interactive HTML5 hierarchical graph of the bacterial diversity in the sample. In this poster we present an easy to use web application tool for the analysis of 16S rRNA fragment and whole sequence data.

Genome dynamics during long-term colonization with KPC+ Klebsiella pneumoniae

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.17

Sean Conlan[1], Clayton Deming[2], James Mullikin[3], Pamela Thomas[3], Morgan Park[3], Karen Frank[4], Tara Palamore[4], Julia Segre[2]

[1]Human Genome Research Institute National Institutes of Health, [2]National Human Genome Research Institute, [3]National Institutes of Health Intramural Sequencing Center (NISC), [4]National Institutes of Health Clinical Center

Long-term colonization with antibiotic-resistant bacteria is a recognized but understudied phenomenon. Recent reports have shown that patients can remain colonized with KPC+ K. pneumoniae for months to years after the detection of a positive culture. In most cases, carriage is tracked by culture-based techniques or PCR and, while these strategies are useful for measuring long-term carriage, they ignore the complex underlying biology of the microbe(s). In particular, continuous colonization can't be distinguished by standard microbiological methods from multiple episodes of colonization or mixed populations within an individual. Furthermore, genomic changes like gain or loss of plasmids are not appreciated. In 2011, the National Institutes of Health (NIH) Clinical Center reported a cluster of KPC+ K. pneumoniae and genomic sequencing was used to impute the transmission map for this clonal cluster. Further analysis using single-molecule real-time sequencing determined that three plasmids were transmitted vertically and maintained over the duration of the outbreak, including the KPC+ plasmid pKpQIL. Persistent carriage was detected in two individuals from the 2011 cluster. In 2013, a KPC+ organism was isolated from patient 15. That isolate was determined to be an ST258 K. pneumoniae, consistent with long term-carriage (21 mo) of the outbreak strain. Alignment of the chromosome to the outbreak strain identified four additional SNVs had accumulated. All three plasmids were detected in the shotgun sequencing but only pKpQIL appeared to be complete. Both pAAC-154 and pKPN-498 have undergone significant rearrangement, resulting in the loss or disruption of genes on each plasmid. Patient 16 has had continuous colonization with KPC+ organisms detectable across multiple timepoints spanning 2011-2015. PCR characterization and whole genome sequencing was used to define a complex pattern of succession and plasmid transmission across two different K. pneumoniae sequence types and an E. coli isolate. These findings demonstrate the utility of genomic methods for understanding strain succession, genome plasticity and long-term carriage of antibiotic-resistant organisms.

Whole Genome Sequencing of Recurrent Methicillin-Resistant Staphylococcus aureus Enables High Resolution Genotyping

Wednesday, 27th May 20.00 - Mezzanine (Sponsored by Roche) – Poster 2b.18

Murtada Alsaadi[1]***, Kimberly Paffett***[2]***, Walter Dehority***[2]***, Jon Femling***[3]***, Renee-Claude Mercier***[4]***, Darrell Dinwiddie***[2]

[1]***School of Medicine, University of New Mexico Health Sciences Center,*** [2]***Department of Pediatrics, University of New Mexico Health Sciences Center,*** [3]***Department of Emergency Medicine, University of New Mexico Health Sciences Center***

Over six million people are naturally colonized with methicillin-resistant Staphylococcus aureus (MRSA) and an estimated 75,000 active infections occur each year in the United States causing significant morbidity, mortality, and financial burden on the healthcare system. Recurrent infections account for up to forty percent of MRSA cases. Determining the genetic composition and evolution of recurrent MRSA may provide critical information to aid in its clinical management and prevention. To this end, we undertook shotgun whole genome sequencing (WGS) of 16 samples from 6 patients with recurrent MRSA infections of the blood. Sequencing of 2x75bp on the Illumina MiSeq resulted in mean whole genome coverage of 50-80X. Sequences were aligned to three prototype complete genome reference strains NC_007793 (USA 300), NC_002745 (USA 100), and NC_007795 using the CLC Bio Genomics Workbench. Whole genome sequencing-based strain typing revealed a 25% discordance rate with previous pulse-filled gel electrophoresis (PFGE) typing. Subsequent variant analysis of recurrent cases uncovered genetic variation in strains isolated from the secondary infection, which may represent evolution of clonal MRSA. Our results indicate that WGS provides higher resolution in identification of MRSA strains as compared to current standard of care PFGE. Furthermore, WGS enables single nucleotide variant detection that can provide insight into the evolution, adaptation, and acquisition of mutations in MRSA during the course of clinical infection. Together, this genomic data has the potential to improve clinical management of pathogenic infections.

Thursday, May 28th Agenda

07:30 - 08:30	*La Fonda Breakfast Buffet*	*Sponsor: NEB*
08:30 - 08:45	Welcome Introduction & Opening Remarks	
08:45 - 09:30	Keynote Address: In Search of the Perfect Assembly (Dr. Daniel Rokhsar)	Sponsor: ThermoFisher
09:30 - 10:30	**Oral Session 3 Genome Assembly & Analysis (Chair: Alla Lapidus & Patrick Chain)**	
09:30 – 09:30	A strategy for creating high quality genome assemblies	David Jaffe
09:50 – 09:50	PBHoney - Detecting SVs with Long-Read Sequencing	Adam English
10:10 – 10:30	De novo assembly of highly repetitive genomes using noisy long-reads	Martin Pippel
10:30 - 11:00	*Coffee Break*	*Sponsor: BioNano Genomics*
11:00 - 12:20	**Oral Session 4 Genome Assembly & Analysis (Chair: Mike Fitzgerald & Bob Fulton)**	
11:00 – 11:20	Chromosome-scale assembly of red raspberry with chromatin interaction libraries	Judson Ward
11:20 – 11:40	Emerging Sequencing, Mapping and Assembly Technologies Enhance Biological Interpretability in the Model Legume Plant, Medicago truncatula	Joann Mudge
11:40 – 12:00	Targeted Genomic Reference Sequences: BAC Sequencing Rejuvenated	Kevin Fengler
12:00 – 12:20	Genome Assembly using Nanopore-guided Long and Error-free DNA Reads	Mohammed-Madoui
12:20 - 13:50	*New Mexican Lunch Buffet*	*Sponsor: Promega*
13:50 - 15:30	**Oral Session 5 Genome Assembly & Analysis (Chair: Darren Grafham & Donna Muzny)**	
13:50 – 14:10	Creating a Platinum Human Genome Assembly	Tina Lindsay
14:10 – 14:30	Using individual scale de novo assembly to identify mutations causing a phenotypic trait	Scott Geib
14:30 – 14:50	Sequencing contamination in genome assemblies as a source of error in bacterial comparative genomics	Jason Sahl
14:50 – 15:10	Pilon: Comprehensive Microbial Variant Detection and Genome Assembly Improvement	Bruce Walker
15:10 – 15:30	Expanding the SPAdes Toolbox	Anton Korobeynikov
15:30 - 15:45	*Coffee Break*	*Sponsor: Lucigen*
15:45 - 17:30	**Tech Talks: Assembly & Analysis (Chair: Johar Ali & Mike Fitzgerald)**	
15:45 – 16:00	PacBio read error correction in the new CLC Genome Finishing Module, a QIAGEN bioinformatics solution for CLC Workbenches enabling non-bioinformatics experts to finish genomes	Andreas Sand
16:00 – 16:15	Whole read overlap assembly accurately detects structural variants now in GRCh38	Niranjan Shekar
16:15 – 16:30	The Road to "Platinum Genomes" – Integrating Genome Mapping in Nanochannel Arrays and NGS Sequencing for Cost-effective Reference Quality Genome Assembly	Palak Sheth
16:30 – 16:45	Geneious R8: a bioinformatics platform for biologists	Christian Olsen
16:45 – 17:00	Integrated Genome Mapping in Nanochannel Arrays and Sequencing for Better Human Genome Assembly and Structural Variation Detection	Andy Wing Chun Pang
17:00 – 17:15	Improving do novo genome assemblies with in vitro chromatin proximity data	Nicholas Putnam
17:15 – 17:15	De Novo Diploid Genome Assembly and Haplotype Sequence Reconstruction	Jason Chin
18:00 - 20:00	***Happy Hour(s) Cowgirl Cafe***	*Sponsor: Illumina*

In Search of the Perfect Assembly

Thursday, 28th May 8.45 - La Fonda Ballroom (Sponsored by ThermoFisher) - Keynote/Plenary

Daniel Rokhsar
University of California Berkley and the Joint Genome Institute

Reconstructing a complete, accurate, phased, chromosome-scale genome sequence from cheap and simple inputs is a holy grail of genomics. This presentation will discuss progress and challenges in genome assembly drawing examples from a diverse array plant and animal genomes of varying complexity and scale.

A strategy for creating high quality genome assemblies

Thursday, 28th May 9.30 - La Fonda Ballroom - Oral

David Jaffe[1], Neil Weisenfeld[1]
[1]*The Broad Institute*

Historically, genome reference sequences have been the product of expensive, painstaking and often ad hoc projects. Typically homologous chromosomes have been collapsed (through inbreeding or computational methods). Because a loss of heterozygosity is generally deleterious, biological inferences based on single-chromosome analysis can be misleading.

Here we propose an affordable method for generating phased genome reference sequences. Our method begins with a microgram of DNA. We make two libraries. The first library is made from 0.5 kb fragments without PCR. The second library is made from 20-200 kb starting DNA molecules, which are compartmentalized and from which barcoded libraries are created using a technology developed by 10X Genomics. Both libraries are deeply sequenced using Illumina sequencers.

Once the data are generated, we create an assembly graph from data of the first type using DISCOVAR de novo, then use the second data type to walk through the graph, so as to represent parts of individual (phased) chromosomes.

We describe assemblies of the first data type (PCR-free). For example for human genomes, these assemblies have contigs of size 100 kb+. Then we describe our first steps in implementing use of the second data type (10X). Quite difficult regions can be resolved, including some segmental duplications that have thus far not been assembled using either short or long WGS reads.

PBHoney - Detecting SVs with Long-Read Sequencing

Thursday, 28th May 9.50 - La Fonda Ballroom - Oral

Adam English[1]***, William Salerno***[1]***, Eric Boerwinkle***[1]***, Richard Gibbs***[1]
[1]***Baylor College of Medicine***

Long-read sequencing (>1 kbp) offers more complete genomic information than does short-read sequencing (~100 bp), but the relatively high cost-per-base limits the practicality of long reads as the sole data source in high-throughput whole-genome sequencing projects. An alternate, more cost-effective strategy is to supplement short-read data with long-read sequencing, which has been effectively implemented by de novo assembly tools including pacbioToCA and PBJelly. Here we describe the performance of our long-read structural variation detection software PBHoney (PMID: 24915764) in human haploid, diploid, and targeted sequencing experiments and compare these results to short-read methodologies.

PBHoney makes 9,496 SV calls between 50 bp and 100 kbp from 40x PacBio coverage of the haploid cell-line CHM1-tert. We then run Parliament, a consolidation SV discovery tool, on 134x Illumina data to generate ~25,000 variant loci. These results are compared to an independently derived set of 12,047 SV calls from the same PacBio data (Chaisson et al. PMID: 25383537). This intersection shows 71% of PBHoney and 56% of Chaisson calls are either matched (5.8k variants) or unmatched between sets. However, this comparison also identifies ~2.5k distinct genomic loci containing ~3.4k ambiguously matched calls per set, which represent uncertainty in the number and nature of structural variants we expect to observe with long-read sequencing.
We next explore PBHoney's efficacy when applied to a diploid human sample that has been previously characterized via multiple technologies including aCGH and BioNano Irys optical mapping. Using 20x coverage (i.e., 10x per haploid genome), we identify 9k deletion, 15k insertion, and 338 inversion calls. Finally, we apply PBHoney to data from targeted sequencing of long-library inserts captured by a novel oligonucleotide enrichment protocol, from which we are able to identify PCR-validated breakpoints within LCR repeats.

Our method is shown to be accurate and fast over a broad range of coverages, size regimes, and SV types. Using long-read sequencing as either the sole or a complementary data source for a SV detection study provides a comprehensive assessment of genomic architecture. By comparing PBHoney's results with SVs identified from Illumina short-reads and other methodologies, we illustrate the additive value of PBHoney and PacBio data.

De novo assembly of highly repetitive genomes using noisy long-reads

Thursday, 28th May 10.10 - La Fonda Ballroom - Oral

***Martin Pippel*[1], *Siegfried Schloissnig*[1]**
[1]*Heidelberg Institute for Theoretical Studies*

De novo assembly of eukaryotic genomes is still a challenging task. The emergence of single molecule real time sequencing [1], makes it possible to produce reads of >10Kb with near uniform sampling of the genome and randomly distributed sequencing errors. Thereby, allowing the accurate reconstruction of almost any genome. In reality, however, the numerous repeat elements and the leniency of the alignments, necessitated by the noise levels of the reads, resulted in massive requirements towards compute time, storage (of local alignments) and subsequent processing in the assembly pipeline.

Building on the Dazzler platform [2], we present two novel approaches for the de novo assembly of highly repetitive genomes in the gigabase range.

In order to reduce the demands towards compute time and storage, we developed a method of dynamically identifying and excluding repeat elements from the overlapping phase, effectively eliminating a large fraction of the calculated local alignments.

Current long-read assemblers [3-6] correct noisy reads prior to assembly, effectively overlapping them twice in the whole process. We find that further increases in efficiency can be achieved by skipping correction altogether and directly assembling the noisy long-reads.

Combined, these two enhancements result in decreases in compute time and storage demands of one order of magnitude and make assemblies of complex genomes on cluster environments of modest size possible.

[1] Korlach J, et al., Chapter 20 – Real-Time DNA Sequencing from Single Polymerase Molecules
Methods in Enzymology, 2010, 472, 431-455
[2] Myers E W, Efficient Alignment Discovery amongst Noisy Long Reads
WABI 2014 - 14th Workshop on Algorithms in Bioinformatics, Wroclaw, Poland, 2014
[3] Myers E W, et al. A whole-genome assembly of Drosophila.
Science, 2000, 287, 2196-2204
[4] Gnerre S, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data
Proceedings of the National Academy of Sciences USA, 2011, 108, 1513-1518
[5] Bankevich A, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology, 2012, 19, 455-477
[6] C. S. Chin, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data
Nature Methods, 2013, 10, 563-569

Chromosome-scale assembly of red raspberry with chromatin interaction libraries

Thursday, 28th May 11.00 - La Fonda Ballroom - Oral

Judson Ward[1], Jasbir Bhangoo[1], Michael Alonge[1]
[1]Driscoll's

The highly heterozygous genome of a red raspberry (Rubus idaeus) cultivar was assembled using ALLPATHS resulting in a draft genome containing 7,332 scaffolds with an N50 scaffold length of 302,924 nt and a total length of 309,259,397nt (including gaps). Chromatin interaction data (HiC) was generated in order to produce long-range scaffolding information for use in order and orienting with LACHESIS. A parameter bracketing experiment was performed using LACHESIS and the impacts on assembly errors and error type associated with parameter changes were identified. Specific errors resulting from heterozygosity were examined and assembly improvements made based on these results. The final, chromosome-scale assemblies were validated through comparison to a dense genetic map and through examination of synteny with high quality assemblies in the rose family. This genome will support traditional plant breeding efforts to produce flavorful and nutritious berries and will also help to inform evolutionary studies in Rubus and across the Rosaceae. Computational methods will inform assembly strategies using chromatin interaction libraries in other complex heterozygous genomes.

Emerging Sequencing, Mapping and Assembly Technologies Enhance Biological Interpretability in the Model Legume Plant, Medicago truncatula

Thursday, 28th May 11.20 - La Fonda Ballroom - Oral

***Joann Mudge*[1], *Thiruvarangan Ramaraj*[1], *Brian Walenz*[2], *Li Song*[3], *Peng Zhou*[4], *Peter Tiffin*[4], *Diego Fajardo*[1], *Kevin Silverstein*[4], *Nevin Young*[4], *Jason Miller*[2]**

[1]*National Center for Genome Resources*, [2]*JCVI*, [3]*Johns Hopkins University*, [4]*University of Minnesota*

The plant, Medicago truncatula, is a model for legume and symbiotic genomics. We have resequenced 300 lines and are currently creating de novo assemblies for 25 lines. Most of these lines have been sequenced and assembled with short read technology (Illumina assembled with ALLPATHS-LG). Three of the lines will have additional data available in order to bring them to a higher quality draft assembly. Two of these lines have 20X PacBio in addition to Illumina data and have been assembled using our hybrid assembly pipeline, ALPACA, which combines ALLPATHS-LG, ECtools, and Celera Assembler. We are putting additional effort into the R108 accession, which is important to the community as the source of functional genetics resources but which is phylogenetically very divergent from the A17 reference sequence. We are adding 100X PacBio sequence which will allow a de novo PacBio assembly. In addition, we have a BioNano optical map in hand and a Dovetail Chicago library is underway. Compared to the Illumina-only assemblies, these high-quality assemblies capture a greater degree of chromosomal re-arrangement, gene family expansion, and tandem duplication events from M. truncatula evolution.

Targeted Genomic Reference Sequences: BAC Sequencing Rejuvenated

Thursday, 28th May 11.40 - La Fonda Ballroom - Oral

***Kevin Fengler*[1], Victor Llaca[1], Yun Zhang[1], Stéphane Deschamps[1], Matt King[1], Greg May[1]**
[1]Dupont*

Why sequence and assemble an entire genome when you only care about a few regions? What if a closely related reference genome sequence does not adequately represent the sequence diversity underlying a trait of interest? BAC sequencing is a classical approach for generating assemblies for targeted regions in the absence of a reference genome sequence or in cases where larger structural variations (i.e. CNVs, PAVs, and large INDELS) cannot be teased apart by re-sequencing methods. New tools in the genomics toolbox that 1) facilitate rapid BAC screening (BAC superpool sequencing), 2) generate robust assemblies (single molecule sequencing technology), and that 3) provide comprehensive views of the genomic architecture (single molecule genome mapping) are enabling the creation of targeted reference sequences for any region of interest (gene, locus, QTL, etc.). Deconvolution by Sequencing (DbS) is a method for assigning NGS short reads from sequenced BAC superpools to individual BACs which provides a means for in silico BAC tiling path selection. Selected BACs can be pooled together without barcoding, sequenced, and assembled into high-quality contigs. And for the finishing touch, building a genome map provides a mechanism for validating BAC assemblies in the targeted region, but also for characterizing potential gaps. In an era where rapidly evolving genomics technologies are making whole-genome assembly increasing fashionable, these same technologies are making localized assembly for key regions via BAC sequencing often, a more efficient option.

Genome Assembly using Nanopore-guided Long and Error-free DNA Reads

Thursday, 28th May 12.00 - La Fonda Ballroom - Oral

Mohammed-Amin Madoui[1]***, Stefan Engelen***[1]***, Corinne Cruaud***[1]***, Caroline Belser***[1]***, Laurie Bertrand***[1]***, Adriana Alberti***[1]***, Arnaud Lemainque***[1]***, Patrick Wincker***[1]***, Jean-Marc Aury***[1]

[1]***CEA-Genoscope***

Long-read sequencing technologies were launched a few years ago, and in contrast with short read sequencing technologies, they offered a promise of solving assembly problems for large and complex genomes. Moreover by providing long-range information, it could also solve haplotype phasing. However, existing long-read technologies still have several limitations that complicate their use for most research laboratories, as well as in large and/or complex genome projects. In 2014, Oxford Nanopore released the MinION® device, a small and low-cost single-molecule nanopore sequencer, which offers the possibility of sequencing long DNA fragments.

The assembly of long reads generated using the Oxford Nanopore MinION® instrument is challenging as existing assemblers were not implemented to deal with long reads exhibiting close to 30% of errors. Here, we presented a hybrid approach developed to take advantage of data generated using MinION® device. We sequenced a well-known bacterium, Acinetobacter baylyi ADP1 and applied our method to obtain a highly contiguous (one single contig) and accurate genome assembly even in repetitive regions, in contrast to an Illumina-only assembly.. Our hybrid strategy was able to generate NaS (Nanopore Synthetic-long) reads up to 60 kb that aligned entirely and with no error to the reference genome and that spanned highly conserved repetitive regions. The average accuracy of NaS reads reached 99.99% without losing the initial size of the input MinION® reads.

We described NaS tool, a hybrid approach allowing the sequencing of microbial genomes using the MinION® device. Our method, based ideally on 20x and 50x of NaS and Illumina reads respectively, provides an efficient and cost-effective way of sequencing microbial or small eukaryotic genomes in a very short time even in small facilities. Moreover, we demonstrated that although the Oxford Nanopore technology is a relatively new sequencing technology, currently with a high error rate, it is already useful in the generation of high-quality genome assemblies.

Creating a Platinum Human Genome Assembly

Thursday, 28th May 13.50 - La Fonda Ballroom - Oral

***Tina Lindsay*[1], *Bob Fulton*[1], *Karyn Meltz Steinberg*[1], *Richard K. Wilson*[1]**

[1]*The Genome Institute at Washington University in St. Louis*

The human genome reference sequence has provided a foundation for studies of genome structure, human variation, evolutionary biology and human disease. At the time the reference genome was originally completed, it was clear, that there were some loci recalcitrant to closure with the technology and resources. What was not clear was the degree to which structural variation and diversity affected our ability to produce a representative genome sequence at these loci. Many of these regions in the genome are associated with large, repetitive sequences. They exhibit complex allelic diversity such that de-convoluting these regions with DNA from a single donor can be complicated. In order to eliminate the complications of multiple alleles, we have utilized DNA from a hydatidiform mole, which is essentially haploid. The first hydatidiform mole sample to be sequenced was CHM1, which has been sequenced and assembled using both Illumina and Pacific Biosciences data. We now have a second hydatidiform mole sample that has been sequenced in the same manner, CHM13. Both of these genomes have been sequenced to ~60X depth of coverage with PacBio data, and have BAC libraries that can be utilized to assemble difficult regions. The goal for both of these genomes is to produce a Platinum, reference-grade assembly, that could be used as a single haplotype reference. In this presentation I will compare the assemblies and sequencing methods used to create both of these assemblies, as well as talk about the methods planned to bring these assemblies to a Platinum status.

Using individual scale de novo assembly to identify mutations causing a phenotypic trait

Thursday, 28th May 14.10 - La Fonda Ballroom - Oral

Scott Geib[1], Sheina Sim[2], Bernarda Calla[1]
[1]USDA-ARS Daniel K. Inouye US PBARC, [2]University of Hawaii, Manoa

The Mediterranean fruit fly (medfly) is an important agricultural pest of many fruit and vegetable species. To protect the mainland United States from this pest, the sterile insect technique (SIT) is employed, involving release of tens of millions of sterile male medfly into the Los Angeles basin of California weekly. These flies have several mutations making them amenable to mass release. Using a chromosomal translocation between the 5th chromosome and the male Y chromosome, females are homozygous recessive for both a temperature sensitive lethal mutation and a white pupal mutation, allowing straightforward separation of male and female flies and generation of male only release strains. Males are maintained heterozygous for these alleles through the chromosomal translocation, linking wild-type phenotype with the sex chromosome. While the relative position of these mutations is known (5th chromosome), the genes and specific mutation causing the traits are not known. To address this, we developed a crossing scheme to isolate these mutations from the SIT line in the background of an inbred lab line (see Sim et al submission). This cross allowed QTL analysis to identify regions of the genome that are strongly linked to these traits. To identify causative mutations, we performed whole genome sequencing of individual F4 flies resulting from this cross including individuals that had the observed mutated phenotypes and those that did not. Utilizing 2 X 250 bp paired-end sequencing on Illumina HiSeq 2500 and subsequent assembly and analysis utilizing DISCOVAR/DISCOVAR-denovo, we were able to generate individual scale assemblies and compare these assembly graph structures surrounding the QTL locations to identify specific loci and mutations in the genome that are consistent with the phenotypes observed in the individuals sequenced. This list represents potential causative mutations for the traits in this SIT line, and currently confirmation of the causative mutations is be determined using targeted gene editing approaches with CRISPR/CAS9, to recreate these mutations in wild-type lab lines. This experiment demonstrates the utility of comparing individual-scale genomic assemblies in non-model organisms to reveal direct structural variations between these assemblies, in contrast to relying on reference based mapping approaches (e.g. GATK) to identify variants between individuals.

Sequencing contamination in genome assemblies as a source of error in bacterial comparative genomics

Thursday, 28th May 14.30 - La Fonda Ballroom - Oral

Jason Sahl[1], Jim Schupp[1], Paul Keim[2]
[1]Translational Genomics Research Institute, [3]Northern Arizona University

Genome assembly is a standard protocol in many comparative genomics workflows and applications. For high throughput bacterial genomics on the Illumina platform, tens to hundreds of genomes are typically multiplexed and sequenced on a single lane or flowcell, with each sample containing a unique index or barcode. Reads are then de-multiplexed into discrete samples and assembled separately. New assembly algorithms, such as SPAdes, accommodate uneven coverage across assemblies and keep as many contiguous regions as possible in spite of potential coverage anomalies. However, during the sample processing, we have observed a small amount of sequence (~1%) that is clustered and associated with the incorrect index during de-multiplexing; we term this "optical contamination". While this level of contamination often doesn't interfere with downstream applications, we demonstrate two cases where low-level optical contamination has led to incorrect biological conclusions based on comparative genomics. We also describe how our assembly pipeline, the universal genome assembly pipeline (UGAP), helps in the identification of problematic contigs from SPAdes assemblies. UGAP represents a high-throughput solution to bacterial genome assembly and is capable of assembling hundreds of bacterial genomes in a matter of hours on a high performance computing (HPC) cluster. We also demonstrate how using this pipeline in conjunction with a dual-indexing approach can drastically reduce the optical contamination problem. Methods are included with UGAP to retrospectively check for anomalies in past genome assemblies from any genome assembler that outputs a multi-FASTA file.

Pilon: Comprehensive Microbial Variant Detection and Genome Assembly Improvement

Thursday, 28th May 14.50 - La Fonda Ballroom - Oral

Bruce Walker[1], Thomas Abeel[2], Terrance Shea[3], Sarah Young[3], Ashlee Earl[3]
[1]Applied Invention, LLC, and the Broad Institute, [2]Delft University of Technology and The Broad Institute, [3]The Broad Institute

We present Pilon[1], a fully automated, all-in-one tool for correcting draft assemblies and calling sequence variants of multiple sizes, including very large insertions and deletions. Conceptually, Pilon treats assembly improvement and variant detection as the same process. Both start with an input genome -- either an existing draft assembly or a reference assembly from another strain -- and use evidence from read alignments to identify specific differences from the input genome supported by the sequencing data. Applying those changes to a draft genome assembly yields an improved assembly, while reporting the changes with respect to a reference genome yields variant calls.

In genomic regions where read alignments are poor, Pilon is capable of filling out and correcting sequence through an internal local reassembly process. This capability allows Pilon to further improve assemblies by filling gaps and correcting local mis-assemblies, and it also enables Pilon to capture many large insertion, deletion, and block substitution variants in their entirety. These larger events are often completely missed or inaccurately characterized by conventional variant calling tools that rely solely on read alignments. Pilon has built-in heuristics to determine which corrections and calls are of high confidence, so no separate filtering criteria need be specified. This allows for the automated processing of hundreds or thousands of data sets representing different microbial species with minimal human intervention.

We have benchmarked Pilon both as an assembly refinement tool and variant caller. Pilon-improved assemblies were more contiguous and complete than non-Pilon-improved assemblies. For variant calling, Pilon performed as well or better when compared with two state-of-the-art variant detection tools in calling small variants, and Pilon differentiated itself in its ability to identify large-scale variants.

Pilon has been used in production at the Broad institute to automatically improve the quality of over 8,000 prokaryote and eukaryote genome assemblies, ranging in size from bacterial to insect, prior to their submission to Genbank, and it has been used to call variants on over 6,000 bacterial and fungal strains.

We will also present results obtained from applying Pilon to large collections of Mycobacterium tuberculosis (Mtb) strains as part of the Tuberculosis Antibiotic Resistance Catalog (TB-ARC) project that have enabled analyses of Mtb phylogeography, transmission, and the molecular mechanisms driving antibiotic resistance.
Pilon is freely available as open source software (https://github.com/broadinstitute/pilon) under the GPLv2 license.

[1] Bruce J. Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K. Young, Ashlee M. Earl (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE 9(11): e112963. doi:10.1371/journal.pone.0112963

Expanding the SPAdes Toolbox

Thursday, 28th May 15.10 - La Fonda Ballroom - Oral

Anton Korobeynikov[1], Dmitry Antipov[1], Anton Bankevich[1], Elena Bushmanova[1], Alexey Gurevich[1], Alla Lapidus[1], Sergey Nurk[1], Andrey Prjibelski[1], Yana Safonova[1], Pavel Pevzner[2]
[1]Saint Petersburg State University, [2]University of California San Diego

Despite its central role in genomics, genome assembly continues to present a challenge. Moreover, the proliferation of new sequencing technologies and the new types of genomic libraries introduce additional levels of complications.

Our Saint Petersburg Assembler (SPAdes) was originally developed for the purpose of overcoming the complications associated with single-cell MDA amplified microbial data (uneven coverage, increased level of errors and chimerical reads). The tool was able to successfully resolve these issues for Illumina reads and was recognized by the scientific community as one of the best assemblers working with both isolates and single-cell data.

From very early on SPAdes was designed as a scalable platform, whose capacity could be increased by the addition of complimentary tools on top of it’s existing codebase. We present several novel tools built using the SPAdes technologies that further extend the applicability of SPAdes to new types of data.

PacBio read error correction in the new CLC Genome Finishing Module, a QIAGEN bioinformatics solution for CLC Workbenches enabling non-bioinformatics experts to finish genomes

Thursday, 28th May 15.45 - La Fonda Ballroom - Tech Talk

Andreas Sand[1], Leif Schauser[1], Aske Simon Christensen[1], Martin Bundgaard[1], Arne Materna[1]
[1]QIAGEN Aarhus A/S

SMRT sequencing technology, as implemented by Pacific Biosciences PacBio), has shown to greatly improve the completeness of genome sequence assemblies, as the read length for this data type is greater than the length of most genomic repeats.

A major obstacle for most assemblers is the high (10-15%) error rate of PacBio reads. Advantages intrinsic to this data type include the randomness of most sequencing errors, and the fact that reads are randomly sampled across the genome. These advantages can be exploited to correct errors in long PacBio reads, as error-corrected long reads are more amenable to genome assembly.

A second obstacle is the presence of chimeric reads as well as sequences derived from adapters, which can be hard to recognize given the error profile of the raw reads.

We have implemented a novel software tool, the CLC Genome Finishing Module (GFM) for CLC Workbenches (now part of the QIAGEN Bioinformatics product portfolio) allowing scientists to make use of PacBio reads through the familiar graphical user interface of the widely used CLC Genomics Workbench.

The tool automates the bioinformatics operations required for data analysis: i) error correction ii) de novo genome assembly using PacBio reads, or via hybrid (datasets include a mix of PacBio and short reads) approaches. In contrast to HGAP, which performs the string-overlap graph based assembly as implemented in the Celera assembler, the CLC GFM makes use of a modified De Bruijn graph based assembly. This approach vastly improves performance and decreases memory requirements, thereby enabling the assembly of microbial genomes from PacBio reads in less than half an hour on a laptop computer.

In terms of assembly quality, CLC GFM performs favorable in benchmarks against the commonly used open source software HGAP.

Whole read overlap assembly accurately detects structural variants now in GRCh38

Thursday, 28th May 16.00 - La Fonda Ballroom - Tech Talk

Becky Drees[1], Noah Spies[2], Jeremy Bruestle[1], Niranjan Shekar[1]
[1]Spiral Genetics, [2]Stanford University School of Medicine& National Institute of Standards and Technology Genome in a Bottle

Variant calling can be used to improve references. However, to do so requires the ability to detect and report the sequence of long insertions. It also requires the detection of variants with a low false discovery rate and with little imprecision. Until now, variant calls for structural variants (SVs) have been prone to imprecision and high rates of false discovery. Here we show how read overlap assembly to detect variants (Anchored Assembly) has the ability to detect indels and structural variants (including larger insertions) with considerable precision and few false discoveries.

In simulated data and in a previous study using real data by English et al (2015, accepted), the method has been shown to have high sensitivity compared to other bioinformatic SV callers and a false discovery rate of less than 5% when detecting SVs.

Using the Anchored Assembly method on Illumina HiSeq data sets of the individual NA12878, we detect variants previously reported in the results of the 1000 genomes project, as well as variants that have now been incorporated into the new human reference (GRCh38). In addition, we detect a number of novel SVs, which were confirmed using PacBio sequencing of fosmids (Eichler et al, 2015, in preparation). This confirms that the method can be used to detect variants not otherwise normally detected by other bioinformatic methods on Illumina data.

Using Anchored Assembly on the Ashkenazi Jewish trio sequenced by Personal Genome Project (PGP), We used the tool SVViz to validate the structural variants by the similarity and segregation within the family. Of 10 SVs in the offspring selected at random, all showed logical consistency of calls and segregation within the trio. Specifically, Anchored Assembly detected a 3.4kb insertion inherited in the offspring that is a match to an alternate allele assembly now in GRCh38. Such was the resolution of this insertion, it was possible to identify the 5 SNPs and an indel within the insertion that were inherited from the father from a single SNP inherited from the mother. This level of precision makes Anchored Assembly useful for family analyses; specifically, to be able to identify all the variants in a proband, dividing them into those that are unique, or de novo, those that are inherited from each of the parents, and those that are common to all family members.

Whole read overlap assembly allows for accurate and precise calls of structural variant with low false discovery rates. The ability to compare across samples is particularly useful for identifying variants associated with autism and rare genetic disorders, as well as improving reference genomes and investigating differences across microbe and plant individuals.

The Road to "Platinum Genomes" – Integrating Genome Mapping in Nanochannel Arrays and NGS Sequencing for Cost-effective Reference Quality Genome Assembly

Thursday, 28th May 16.15 - La Fonda Ballroom - Tech Talk

***Palak Sheth*[1], *Andy Wing Chun Pang*[1], *Alex Hastie*[1], *Thomas Anantharaman*[1], *Zhanyang Zhu*[1], *Heng Dai*[2], *Zeljko Dzakula*[1], *Han Cao*[1]**
[1]*BioNano Genomics*, [2]*WuXi AppTec*

High-quality reference genome assembly and finishing continues to be a time-consuming and cost-prohibitive process involving many experts and manual curation. Assemblies based solely on short-read technologies are often fragmented due to structural complexities such as repetitive regions, long-range structural variations (SVs), and dispersed segmental duplications. Assemblies integrating multiple data types often require different sample preparations and data collection methods, which, in turn, add additional complexities, time and cost. Sub-optimal quality reference genomes often wrongly associate or do not detect medically relevant genes leading to incomplete SV detection and characterization.

The BioNano Genomics Irys System linearizes long DNA molecules, thus yielding single-molecules preserving long-range information. These hundreds of kilobases molecules can span interspersed repeats and capture structural information that are often missed by other sequencing platforThe assembled genome maps can scaffold sequencing assemblies to validate the accuracy of the sequences, and to anchor the adjacent sequences into the proper order and orientation.

We present a comprehensive sample to answer workflow integrating single-molecule genome mapping assembly with NGS assembly to generate reference quality genome assemblies. We present results of hybrid scaffolds from a wide variety of samples such as human, Arabidopsis, tomato, kingfish, duckweed, and banana, were sequencing and genome mapping technologies correspond well. The resulting hybrid scaffolds are highly contiguous with N50s far exceeding those achieved by NGS sequencing alone. We demonstrate that cost and time-effective near-reference quality genome assembly is now possible with the integration of genome mapping and NGS sequencing. High-quality references generated using this workflow are indispensable for accurate and comprehensive SV detection and biological impact assessment, enabling characterization of complex SVs missed by other technologies.

Geneious R8: a bioinformatics platform for biologists

Thursday, 28th May 16.30 - La Fonda Ballroom - Tech Talk

***Christian Olsen*[1], *Kashef Qaadri*[1], *Richard Moir*[1], *Matt Kearse*[1], *Matthew Cheung*[1], *Jonas Kuhn*[1], *Alex Cooper*[1], *Chris Duran*[1]**

[1]*Biomatters*

Biomatters’ Geneious R8 is a bioinformatics software platform that allows researchers the command of industry-leading algorithms and tools for their genomic and protein sequence analyses. Using a glass-box approach for software design, Geneious R8 offers a comprehensive suite of peer-reviewed tools that enable biologists to be more efficient with their bioinformatic workflows. Researchers at all experience levels can easily manage, analyze, and share their sequence data via a single intuitive Java software application.

R8 provides tools for next-generation sequence analysis, sequence alignment, molecular cloning, chromatogram assembly, and phylogenetics. New features for the R8 major version release include a 16S rRNA 'Biodiversity Tool', CRISPR gene editing, circular de novo assembler, improved 'Workflow Builder' as well as a number of new plug-ins. R8 affords real-time dynamic interaction with sequence data and empowers biologists to produce stunning publication quality images to increase the impact of their research. By utilizing Geneious R8, biologists can easily improve their sequence analysis workflow efficiencies to free up more time for their research. This tech talk aims to demonstrate the new features and benefits of the highly integrated Geneious R8 tool-suite.

Integrated Genome Mapping in Nanochannel Arrays and Sequencing for Better Human Genome Assembly and Structural Variation Detection

Thursday, 28th May 16.45 - La Fonda Ballroom - Tech Talk

Andy Wing Chun Pang*[1], Alex Hastie[1], Palak Sheth[1], Thomas Anantharaman[1], Zhanyang Zhu[1], Zeljko Dzakula[1], Heng Dai[2], Han Cao[1]
[1]BioNano Genomics, [2]WuXi AppTec

De novo genome assemblies using purely short sequence reads are generally fragmented due to complexities such as repeats found in most genomes. These characteristics can hinder short-read assemblies and alignments, and that can limit our ability to study genomes.

The BioNano Genomics Irys System linearizes long DNA molecules, thus yielding single-molecules containing long-range information. These hundreds of kilobases molecules can capture structural information that may be missed by other sequencing platforThe assembled genome maps from these molecules can scaffold sequencing assemblies to validate the accuracy of the sequences, and to anchor the adjacent sequences into the proper order and orientation. The long-range hybrid scaffolds are able to identify novel chromosomal rearrangements undetectable by short-read alignment or reference-guided assembly approaches.

We present a comprehensive analysis of a human genome by combining single molecule genome mapping with one of the most annotated sequence assemblies, the HuRef assembly. Overall, we found that the assemblies of sequencing and genome mapping technologies correspond well, and the resulting hybrid scaffolds are highly contiguous, with a N50 exceeding 35Mb, a value typically unachievable by short-read sequencing. In addition, we compared the structural variation with calls previously detected in the HuRef assembly, and found multiple novel variants spanning over hundreds of kilobases in size. Some of these variants reside in areas where the sequence assembly was poorly covered or was highly fragmented; yet these variants encompass numerous genes, and can be of functional importance. Finally, we identified genome maps that span over the remaining reference gaps, and maps that resolve and measure long tandem repeats.

Improving do novo genome assemblies with in vitro chromatin proximity data

Thursday, 28th May 17.00 - La Fonda Ballroom - Tech Talk

***Nicholas Putnam**[1], **Jonathan Stites**[1], **Brandon Rice**[1], **Charles Sugnet**[1], **Brendan O'Connell**[1], **Paul Hartley**[1], **Andrew Fields**[1], **David Haussler**[2], **Daniel Rokhsar**[3], **Richard Green**[2]*

[1]Dovetail Genomics, LLC, [2]University of California Santa Cruz, [3]University of California Berkeley

Proximity ligation methods, including Hi-C and Dovetail Genomics' Chicago library preparation, transform spatial information about DNA molecule conformation into a form that can be read out on the highest-throughput, lowest-cost DNA sequencing platfor Hi-C encodes structural information about chromosomes in vivo, while Chicago libraries are constructed in vitro beginning from naked DNA. We describe the likelihood model and computational methods we have developed for improving draft genome assemblies, and describe the results of their application to vertebrate, invertebrate, and plant genomes. Combined with short fragment paired-end sequencing, a single Chicago library can replace multiple mate-pair and fosmid end libraries while improving both contiguity (N50) and accuracy.

De Novo Diploid Genome Assembly and Haplotype Sequence Reconstruction

Thursday, 28th May 17.15 - La Fonda Ballroom - Tech Talk

***Jason Chin**[1], **Paul Peluso**[1], **David Rank**[1]*
*[1]**Pacific Biosciences***

Polymorphisms between homologous chromosomes pose algorithmic challenges for assembling diploid genomes. In particular, structural polymorphisms may limit the contiguity of genome assemblies. Many genome projects choose inbred lines (or in the case of humans a hydatidiform mole) to emulate haploid genomes to overcome such limitations. Here we present novel strategies and algorithms that are capable of assembling diploid genomes and reconstructing the haplotype sequence of each chromosome using long DNA sequence reads. Long continuous reads and lack of systematic errors are essential for generating high quality assemblies. Having such characteristics, PacBio® SMRT® reads provide crucial information to resolve polymorphisms between homologous chromosomes in diploid genomes. The current graph layout heuristics is already capable of processing the structural level differences and maintaining assembly contiguity in the experimental Falcon assembler (https://github.com/PacificBiosciences/FALCON). Falcon generates long hybrid contigs (primary contigs) from both homologous chromosomes and associated contigs representing structural differences between them. Moreover, one can segregate the reads associated with their different origins by heterozygous SNPs and structural variations. With segregated read groups, not only can one generate phased SNP calls, but the original haplotype sequences including structural variation can also be reconstructed as haplotype contigs. We developed a new tool, "Falcon Unzip," for constructing such haplotype contigs for diploid genome assembly. We first test and evaluate Falcon Unzip on synthetic diploid data. We apply it on a diploid human genome to show its capability on reconstructing haplotype contigs for complicated regions like KIR and MHC regions.

Happy Hour(s) @ Cowgirl Café

Cowgirl

505.982.2565   319 S. Guadalupe St   Santa Fe, NM

See map on next page!

6:00pm – 8:00pm, May 29th

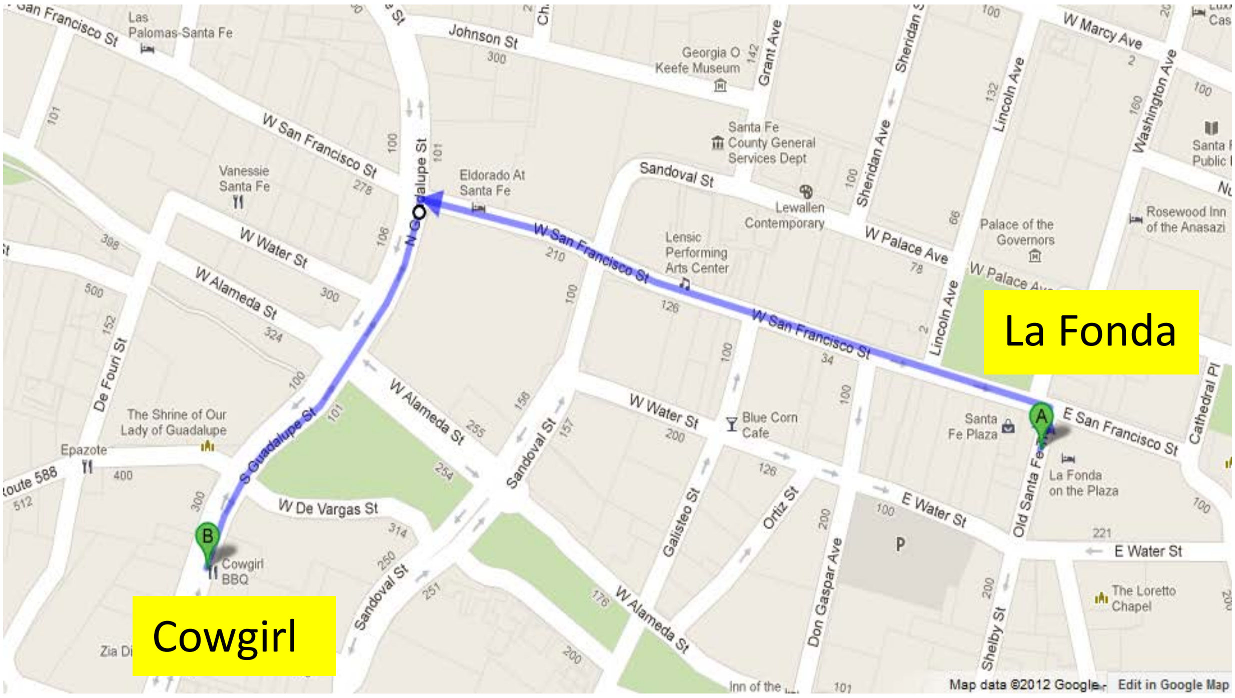
Drink tickets (margaritas, beer, sodas) provided

Sponsored by illumina!!!

Enjoy!!!

Map to Cowgirl

505.982.2565   319 S. Guadalupe St   Santa Fe, NM



Total Walking Distance

0.5 miles, 10 minutes

The Legend...

Many years ago, when the cattle roamed free and Cowpokes and Cowgirls rode the range, a sassy young Cowgirl figured out that she could have as much fun smokin' meats and baking fine confections as she could bustin' broncs and rounding up outlaws. So she pulled into the fine bustling city of Santa Fe and noticed that nobody in town was making Barbeque the way she learned out on the range. She built herself a Texas-style barbecue pit and soon enough the sweet and pungent scent of mesquite smoke was wafting down Guadalupe street and within no time at all folks from far and near were lining up for heaping portions of tender mesquite-smoked brisket, ribs and chicken. Never one to sit on her laurels, our intrepid Cowgirl figured out that all those folks chowing down on her now-famous BBQ need something to wash it all down with. Remembering a long-forgotten recipe from the fabled beaches of Mexico, she began making the now-legendary Frozen Margarita and the rest, as we say, is History. Before you could say "Tequila!" the musicians were out playing on the Cowgirl Patio and the party was in full swing.

Friday, May 29th Agenda

07:30 - 08:30	*Harvey House Breakfast*	*Sponsor: NEB*
08:30 - 08:45	Opening Remarks	
08:45 - 09:30	Keynote Address Evolution and Epidemiology of Anthrax through lens of Genome Analysis. ((Dr. Paul Keim)	Sponsor: Advanced Analytic
09:30 - 10:30	**Oral Session 6 Pathogens & Microbial Genomics (Chair: Donna Muzny & Bob Fulton)**	
09:30 – 09:45	Living on the EDGE: Robust generalized bioinformatics for next-gen sequencing novices	Patrick Chain
09:45 – 10:00	Use of whole genome sequencing to account for drifting genomic profiles in rapidly evolving outbreaks	Sterling Thomas
10:00 – 10:15	The Evolution of Pathogen Discovery in Liberia	Lawrence S. Fakoli III
10:15 – 10:30	Use of next generation sequencing for identification and characterization of pathogens associated with undifferentiated fevers	Beth Mutai
10:30 - 10:45	*Coffee Break*	*Sponsor: PacBio*
10:45 - 12:45	**Oral Session 7 Pathogens & Microbial Genomics** **(Chair: Mike Fitzgerald & Patrick Chain)**	
10:45 – 11:00	Lyve-SET: a high-quality SNP pipeline for aiding in bacterial pathogen outbreak investigation	Lee Katz
11:00 – 11:15	RIGEL: An Integrated System for Rapid Detection and Characterization of Known and Unknown Microbes	Willy Valdivia
11:15 – 11:30	Metagenomic pathogen detection and gut microbiome response to acute Salmonella infection	Andrew Huang
11:30 – 11:45	Metagenomic Profiling and Identification of Antimicrobial Resistance Genes from Indoor and Outdoor Airborne Microbial Communities	Tamar Dickerson
11:45 – 12:00	Study of the genetic traits associated with antibiotic resistance in Staphylococcus aureus isolated from skin wards of KPK, Pakistan	Abid Khan
12:00 – 12:15	Epigenomic characterization of Neisseria gonorrhoeae isogenic mutants and clinical isolates to examine the role of DNA methylation in antimicrobial resistance	A. Jeanine Abrams
12:15 – 12:30	Utility of Ion Torrent PGM Sequencing in Genomic Characterization of Clostridium botulinum	Brian Raphael
12:30 – 12:45	Genetic diversity within the botulinum neurotoxin--producing bacteria and their neurotoxins	Karen Hill
12:45 - 13:30	*Santa Fe Deli Lunch Buffet*	*Sponsor: MRIGlobal*
12:45 - 13:00	CR-1 Closing Remarks (End of Regular Sessions) Wrap up of SFAF2015 and planning for 2016	Chris Detter
13:30 - 15:00	**Forensic Analysis 1 Forensic Applications of NGS** **(Chair: Robert Bull & Kristen McCabe)**	
13:30 – 13:45	Evaluation of Massively Parallel Sequencing Technologies for Expanded DNA Identification Capabilities at the Federal Bureau of Investigation Laboratory	Jodi Irwin
13:45 – 14:00	Validation of a Targeted Next Generation Sequencing Solution for Forensic Genomics	Cydne Holt
14:00 – 14:15	Sequence Diversity within Short Tandem Repeat Loci: Applications to Human Identification	Peter Vallone
14:15 – 14:30	Human Identification Advances Using Next Generation Sequencing Technologies	Joseph Chang
14:30 – 14:45	Assessment of Massively Parallel Sequencing chemistries for forensic casework applications	Lilly Moreno
14:45 – 15:00	A Method for Separating Microbial DNA from Vertebrate DNA	Fiona Stewart
15:00 - 15:15	*Coffee Break*	*Sponsor: Qiagen*
15:15 - 17:00	**Forensic Analysis 2 Forensic Applications of NGS** **(Chair: Robert Bull & Kristen McCabe)**	
15:15 – 15:30	Next-generation Sequencing and Custom Software Analysis of mtDNA Mixtures	Cassandra Calloway
15:30 – 15:45	Evaluation of Concordance and Low Level Variant Detection for Forensic-Quality High-Throughput Sequencing of the Full mtGenome using the Illumina MiSeq	Michelle Peck
15:45 – 16:00	Strengths and Limitations of NGS for Forensic DNA Analysis	Jaynish Patel
16:00 – 16:15	A Universal Microbial Clock for Estimating the Postmortem Interval	Jessica Metcalf
16:15 – 16:30	Microbial Forensics of select agents from trace environmental or clinical samples: making the case for targeted sequencing	Tom Slezak
16:30 – 16:45	Bioforensic Metagenomics at the National Bioforensic Analysis Center	M. J. Rosovitz
16:45 - 17:30	**Round Table 2: Discussion of Forensics Applications for NGS Technologies** **(Chair: Robert Bull)**	

Evolution and Epidemiology of Anthrax through lens of Genome Analysis.

Friday, 29th May 8.45 - La Fonda Ballroom (Sponsored by Advanced Analytic) - Keynote/Plenary

Paul Keim
Northern Arizona University and the Translational Genomics Research Institute

Bacillus anthracis is the causative agent of anthrax and widely feared as a biological weapon. Its ecological and evolutionary biology is driven by its alternating life history as growing vegetative cells and quiescent spore forSpores are stable for long periods of time without growth, which greatly slows its evolution. Its geographic dispersal has occurred for millennia but has accelerated in the last few centuries. Phylogeographic reconstructions suggest that human activities have driven much of its distribution, in recent as well as ancient times. Recent human anthrax has occurred among injectional drug users and whole genome sequences have identified clusters of cases that were not readily apparent using classic epidemiology. Likewise, the use of whole genome sequences has been applied to historical biological weapons production in the Soviet Union.

Living on the EDGE: Robust generalized bioinformatics for next-gen sequencing novices

Friday, 29th May 9.30 - La Fonda Ballroom - Oral

Patrick Chain[1]
[1]***Los Alamos National Laboratory***

With the continuing evolution of sequencing platforms and technologies, the so-called democratization of sequencing is in fact already in full swing. Despite the inherent challenges involved, moving sequencing technology into the field and closer to the source of diverse and interesting biological samples, is an attractive idea to many agencies. This even extends to OCONUS laboratories that are being equipped with next generation sequencing (NGS) platforms to complement more traditional molecular, cell, and microbiology methods for infectious disease research. However, many groups new to NGS and/or laboratories in remote or austere locations may be ill-equipped to handle the bioinformatic requirements associated with rapid production of massive, complex datasets from sequencing clinical or environmental samples. A collaborative effort, named EDGE bioinformatics, has begun to research and prototype bioinformatic pipelines that can be deployed to new NGS laboratories in order to enable successful adoption of sequencing technologies by allowing robust processing of NGS data. These pipelines were initially developed with specific use cases and sample types in mind. The pipelines are being designed to make use of the most common file formats and run in a Linux environment on relatively inexpensive hardware so that barriers to adoption are minimal. A pilot program to install and utilize EDGE at an OCONUS DoD facility has already taken place, with successful processing of locally-generated data using a recently locally-installed MiSeq, and demonstrated reach-back capability using CONUS support.

Use of whole genome sequencing to account for drifting genomic profiles in rapidly evolving outbreaks

Friday, 29th May 9.45 - La Fonda Ballroom - Oral

Sterling Thomas[1]***, Shanmuga Sozhamannan***[2]

[1]***Noblis,*** [2]***Critical Reagents Program***

Genome sequence analysis of the 2014 Zaire Ebola Virus (EBOV) isolates revealed a potential problem with the diagnostic assays currently in use; i.e. drifting genomic profiles of the virus may affect assay sensitivity or produce false-negative results. Using a whole genome approach and BioVelocity we identified new signatures that are unique to each of the EBOV, SUDV, and RESTV. This presentation will focus on two areas: the BioVelocity algorithm and how drifting genetic profiles impacts the response to rapidly evolving outbreaks when using less accurate assays.

The Evolution of Pathogen Discovery in Liberia

Friday, 29th May 10.00 - La Fonda Ballroom - Oral

Lawrence S. Fakoli III[1]

[1]***Liberian Institute for Biomedical Research***

The Liberian Institute for Biomedical Research (LIBR) was founded 40 years ago and at one time was considered the scientific hub of West Africa. Due to a long and brutal fourteen-year civil war, LIBR was left a shell of its former self with close to no funding or resources. In 2010, the Naval Medical Research Unit No. 3 (NAMRU-3) was funded by the Global Emerging Infections Surveillance System (GEIS) to initiate country wide malaria and arbovirus surveillance projects. With minimal funding input and effective guidance by visiting research teams, LIBR's capabilities and research technicians evolved from an under-supported and undertrained laboratory to a highly productive research institute staffed by proficient research technicians. Funding dollars increased equipment and technology capital into LIBR that permitted the development of and training in molecular techniques, such as polymerase chain reaction (PCR) based pathogen identification. At the onset of the West African Ebola virus (EBOV) outbreak beginning in November 2013, the LIBR staff were prepared and trained to assist the National Reference Laboratory with ebola virus disease (EVD) diagnostic testing. The LIBR technical skills and equipment were key in establishing the first EVD diagnostic lab in Liberia. More importantly, the LIBR staff and institute went through a second technology evolution and again adapted to implement advanced genomic capabilities for EBOV genome surveillance. The first 25 EBOV genomes were sequenced and analyzed in-country by LIBR staff, providing genetic resolution of viral mutation that occurred between September 2014 and February 2015. With the likelihood of future re-emergence of EVD in Liberia, the Institute and Staff stand ready to implement diagnostic testing and genomic sequencing to help advice public health measures. The EVD outbreak fostered technological and intellectual advancement at LIBR that is revitalizing and reshaping past malaria and arbovirus surveillance projects with NAMRU-3. Current and future surveillance projects will apply genomic-based pathogen discovery from tick and mosquito collections throughout Liberia to link early-onset of emerging epidemics to point-of collection of malaria- and arbovirus-harboring insects.

Use of next generation sequencing for identification and characterization of pathogens associated with undifferentiated fevers

Friday, 29th May 10.15 - La Fonda Ballroom - Oral

***Beth Mutai*[1], John Waitumbi[1]**

[1]United State Army Medical Research Unit - Kenya / Kenya Medical Research Institute*

Lyve-SET: a high-quality SNP pipeline for aiding in bacterial pathogen outbreak investigation

Friday, 29th May 10.45 - La Fonda Ballroom - Oral

***Lee Katz*[1], Darlene Wagner[2], Aaron Petkau[3], Cameron Sieffert[3], Heather Carleton[2], Shaun Tyler[3], Gary van Domselaar[3]**

[1]Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention, [2]Enteric Disease Laboratory Branch, Centers for Disease Control and Prevention, [3]Public Health Agency of Canada

Modern outbreak investigation is largely enhanced with molecular evidence. These lines of evidence have been, but are not limited to: pulsed-field gel electrophoresis (PFGE), multiple-locus variable number tandem repeat analysis (MLVA), and multilocus sequence typing (MLST). In the age of whole genome sequencing (WGS), outbreak investigation is also being aided by whole genome phylogenetic methods. The basic assumption about using WGS for outbreak investigation is that evolution approximates epidemiology. A corollary to that basic assumption is that a phylogeny will approximate a transmission tree.

Therefore, Lyve-SET was created to uncover high-quality SNPs and create phylogenies for outbreak investigations. The basic steps are 1) map reads to a reference genome; 2) call variants; 3) create a multiple sequence alignment (MSA); and 4) infer a phylogeny. The “high-quality” of hqSNP analysis in Lyve-SET consists of 1) region masking for troublesome loci (e.g., phages); 2) nonambiguous and high-identity (95%) read mapping; 3) a minimum coverage per SNP; 4) a minimum percent consensus per SNP; and 5) maximum likelihood phylogeny. Some parameters can be fine-tuned which might aid investigations on a per-species basis: level of coverage required per SNP, percent consensus required per SNP, allowed spacing between SNPs, and other small tweaks. Lyve-SET also contains modular tools that can also be used separately or even in other pipelines. These tools include, but are not limited to, creating an MSA from variant call format (VCF) files, determining pairwise distances between taxa, applying the fixation index (FST), and removing uninformative sites in an MSA.

There have been many success stories of Lyve-SET including aiding in many Salmonella and Listeria outbreak investigations, cluster investigations, and recalls. Lyve-SET continues to be an aid for outbreak investigations even in the present.

Lyve-SET and its documentation are available at https://github.com/lskatz/lyve-SET.

RIGEL: An Integrated System for Rapid Detection and Characterization of Known and Unknown Microbes

Friday, 29th May 11.00 - La Fonda Ballroom - Oral

Willy Valdivia[1]
[1]Orion Integrated Biosciences Inc.

The threat of terrorist or criminal use of pathogenic organisms and their toxins is a great concern. However, despite the exponential accumulation of microbial genomic information, there is no reference database where researchers can retrieve curated sequences specific to a give taxonomic group. This situation continues to hinder efforts for the rapid development of standardized reagents; biosurveillance; and microbial forensic analyses for attribution. To address these limitations, here we report a genomic and metagenomic analysis system for known and unknown microbe discrimination. This architecture employs new comparative genomics algorithms and large-scale data management system to identify, store and update genomic signatures (GS) and motif fingerprints (MF) specific to a given strain, species, genus or family. With this computational analysis named RIGEL, we determined that approximately 25% of all sequence information available for pathogenic microbes is incorrectly assigned to a given taxonomy and that more than 60% of these data cannot be traced reliably to a specific host or temporal-spatial origin. Our extensive analysis and benchmarking against commonly used genomic and metagenomic analysis algorithms revealed that, depending of the query database [nucleotide (NT), non-redundant (NR), or reference (RefSeq)], the false positive rate or false negative rate of these tools ranges from 5% to 60% in species assignment. For some pathogenic viruses (e.g. Dengue and Foot and Mouth Disease), the error rate in serotype assignment can reach up to 12-25%. RIGEL corrected and disambiguated genomic records and mapped this information to geo-location attributes from the National Geospatial Intelligence Agency. This attribution and forensic system was deployed to discriminate known and known microbes in metagenomics samples. RIGEL exhibits a threefold higher sensitivity in the attribution of complex biological samples. For the recent EBOV outbreak RIGEL identified 76 million Filoviridae DNA reads and 19 million unique genomic segments yielding hundreds of “EBOV assembled lineages” These analyses determined: (1) The microbial taxonomic composition of each sample; (2) The quasispecies distribution of EBOV variants in 298 samples from patients from Guinea and Sierra Leone; (3) The identification of Lassa virus in 12% of the samples, suggesting most likely true co-infection rather than contamination; (4) The mapping of EBOV discharge vs. death patient outcomes to viral read count. RIGEL shows that ZEBOV genomic information in samples collected in Guinea during the early stages of the outbreak were “unique events”. As EBOV spread in Sierra Leone, the virus “reverted” to a more stable variant represented by the isolate Makona-IM095B which occurred in 85% of the samples. Our pipeline for collecting, processing, analyzing and representing complex metagenomic samples from blood, water, soil, feces using different sequencing platforms and the implications of our work in attribution, forensics and the development of field deployable assays will be summarized.

Metagenomic pathogen detection and gut microbiome response to acute Salmonella infection

Friday, 29th May 11.15 - La Fonda Ballroom - Oral

Andrew Huang[1]***, Michael Weigand***[2]***, Angela Pena-Gonzalez***[3]***, Kostantinos Konstantinidis***[2]***, Cheryl Tarr***[1]

[1]***Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention,*** [2]***School of Civil and Environmental Engineering, Georgia Institute of Technology,*** [3]***School of Biology, Georgia Institute of Technology***

Current diagnostic testing for bacterial foodborne pathogens relies on culture-based techniques even though many microorganisms, including known pathogens, cannot be cultured. Powerful sequence-based approaches such as metagenomics have potential to derive epidemiologically-relevant information directly from complex samples, bypassing the need to isolate individual organisHowever, such methods have not been systematically applied to foodborne pathogen detection because standardized bioinformatics techniques for analysis have not been established.

We applied shotgun metagenomics to anonymized residual stool samples collected from foodborne outbreaks attributed to Salmonella to evaluate metagenomics as a diagnostic and disease surveillance tool, as well as to gain insight into the gut microbial community responses to foodborne bacterial infection. These outbreaks were geographically isolated and the etiologic agents were identified by culture methods as distinct strains of Salmonella enterica serovar Heidelberg. We performed shotgun sequencing on these samples using the Illumina MiSeq platform. Community and taxonomic analysis were performed using Parallel-META, Metaphlan, and GOTTCHA. Subspecies analysis was performed using BLAST recruitment analysis. Further phylogenetic analysis was performed on metagenomic assemblies of samples and resulting contigs matching S. enterica.

Sample consistency and human DNA sequence abundance varied greatly, often reducing the sequencing depth of the targeted microbial communities, yet referenced-based detection of Salmonella serovar Heidelberg was possible by metagenomic read recruitment as well as metagenomic assembly, even in samples with high human DNA content (90-96%). Taxonomic profiling revealed similar microbial community structures between individual patients from each localized outbreak; samples from different outbreaks clustered separately and were distinct from a subset of ‘healthy’ references selected from the Human Microbiome Project. Microbial gut communities consistently showed reduced species diversity in each foodborne outbreak compared to ‘healthy’ references.

These results highlight the potential utility of metagenomic-based diagnostic tools for foodborne pathogen identification and epidemiologically relevant clustering, even in samples with high human DNA abundance. Furthermore, shotgun metagenomic approaches offer additional insight into gut microbial community responses to foodborne illness that may hold clues to pathogen ecology.

Metagenomic Profiling and Identification of Antimicrobial Resistance Genes from Indoor and Outdoor Airborne Microbial Communities

Friday, 29th May 11.30 - La Fonda Ballroom - Oral

Tamar Dickerson[1], Michelle Galusha[1], Nicole Waybright[1], Danielle Swales[1], Melissa Krause[1], Jeanette Coffin[1], Peggy Lowary[1], Joseph Bogan[1], Jonathan Jacobs[1]
[1]MRIGlobal

Objective: To assess the temporal dynamics of airborne bacterial communities and the dispersion of AMR genes present within them.

Introduction: Since the adoption of antibiotics in the early 20th century, a plethora of clinical pathogens have acquired resistance to one or more modern-day antibiotics. This has resulted in antimicrobial resistance (AMR) being recognized as a severe threat to human and animal health worldwide. Recent work has demonstrated that AMR bacteria are widely prevalent in the environment, perhaps exacerbated by the widespread use of antibiotics for clinical or agricultural purposes. The extent to which AMR genes are present in airborne microbial communities has largely gone unstudied, despite the role these communities may play as latent reservoirs for acquired resistance.

Methods: Dry air filter units were used to collect air samples daily at four locations inside and outside a transit center in the National Capital Region. Microbial biomass was eluted from each filter, concentrated by ultrafiltration, and DNA was extracted for downstream shotgun metagenomic sequencing on an Ion Proton. In addition, clone libraries were prepared from pooled samples and subjected to a functional metagenomics screen for antibiotic resistance genes against seven antibiotics. Antibiotics tested include chloramphenicol, ciprofloxin, colistin, cefepime, tetracycline, penicillin and meropenem. Confirmation of functional AMR genes is ongoing and includes downstream resequencing and comparisons against the initial shotgun data to assess the prevalence of these genes in the airborne bacterial community.

Results: Initial shotgun metagenomic sequencing data from each of the four sites indicate a diverse population of environmental, human, animal and plant bacteria, with the genus Pseudomonas representing the largest proportion of sequences. Moreover, analysis of the shotgun metagenomics data suggests the presence of various antibiotic resistance mechanisms, whose biological functionality remains to be confirmed by further laboratory testing.

Conclusions: Preliminary results suggest that airborne microbial communities may serve as a dynamic reservoir for the dispersion of antimicrobial resistance factors in the environment, potentially complicating the existing world-wide public health crisis to combat AMR pathogens. Therefore, these results emphasize the importance of instituting a strong National AMR Biosurveillance strategy aimed at identifying novel AMR genes before they are acquired by human bacterial pathogens.

Study of the genetic traits associated with antibiotic resistance in Staphylococcus aureus isolated from skin wards of KPK, Pakistan

Friday, 29th May 11.45 - La Fonda Ballroom - Oral

***Abid Khan*[1], *Ghosia Lutfullah*[2], *Saeed Khattak*[2], *Jehan Bakht*[2], *Sajid Ali*[3], *Ali Johar*[4]**
[1]*COMSATS Institute of Information Technology*, [2]*University of Peshawar*, [3]*Bacha Khan University*, [4]*Alvi-armani*

Objective: Staphylococcus aureus is one of most serious human pathogenic microorganisIt is a major cause of nosocomial and community acquired diseases and can sometimes cause serious and life threatening infections. The infections caused by S. aureus are often difficult to recover easily, due to their antibiotic resistance.

Method: In the present study, PCR technique are used to investigate the genetic traits of resistance in S. aureus, isolated from skin wards of two major hospitals and propose the use of NGS sequencing in the future. A total of 100 samples were collected from both male and female, where 50 were from patient's site of infection and 50 from ward environment.

Results: These results demonstrate that the total prevalence of S. aureus both in ward as well as in patients was 48%. The S. aureus prevalence was the highest in female patients (50%) followed by ward environment (29%) and then male patients (21%). The antibiotic sensitivity tests revealed that the highest (91.6% isolates) sensitivity was shown against Imipenem. However, the highest resistance was found to be against Penicillin (100% isolates) and Cefotaxime (75% isolates). In addition, only 29% of the isolates were found to be resistant to methicillin. PCR technique based on the previously designed primers targeting different genetic traits of resistance revealed that 13 out of 14 isolates resistant to methicillin were positive for mecA gene. blaZ genetic traits were found in all isolates resistant to Penicillin. The multi-drug resistance traits, vgaA and vgaB each were detected only in 12.5% of S. aureus isolates. Hence our findings conclude that the phenotypic character of antibiotic resistance is highly correlated to different genetic traits of resistance.

Conclusions: Based on our findings, it is concluded that antibiotic resistance in S. aureus strains is increasing day by day due to self-medications and medication by non-registered medical practitioners (non-RMP). Therefore, for quick and fast detection, we propose Next-Generation Sequencing (NGS) be utilized to screen for antibiotic resistance.

Epigenomic characterization of Neisseria gonorrhoeae isogenic mutants and clinical isolates to examine the role of DNA methylation in antimicrobial resistance

Friday, 29th May 12.00 - La Fonda Ballroom - Oral

A. Jeanine Abrams[1], Steven Johnson[2], David Trees[2]
[1]Division of STD Prevention, NCHHSTP, Centers for Disease Control and Prevention,
[2]Centers for Disease Control and Prevention, Atlanta, GA

The emergence of multidrug-resistant Neisseria gonorrhoeae has hampered the control and prevention of gonorrhea in the United States and globally. Historically, most antimicrobial resistance in N. gonorrhoeae has resulted from the accumulation of mutations in a variety of chromosomal genes. The presence of a constellation of these mutations results in levels of resistance to a variety of antibiotics that reduce the likelihood of successful therapy, and it has resulted in the general elimination of some antibiotics as useful therapeutic agents. With the appearance of the mosaic form of penA, ceftriaxone MIC values increased 4-10 fold above those previously noted. An apparent consequence of the mosaic form of penA is the occurrence of treatment failures with various cephalosporins, and strains that contain the mosaic form are of greater significance as they are able to mutate to still higher levels of resistance to cefixime, cefpodoxime, and ceftriaxone.

To increase the number of penA mutants available for analysis we developed an approach, described as replicative mutagenesis, which has allowed us to isolate large numbers of mutants in the penA gene. We used this approach to isolate a set of nine mutants in gonococcal strain 3502, which contains a mosaic-type penA gene. The MIC values to ceftriaxone for the 3502APMx mutants are >1.0 µg/mL. These 3502APM mutants can be manipulated to increase their ceftriaxone MIC values to 6.0-8.0 µg/mL (3502APMx-x strains). Thus, the effects of mutations in the mosaic penA can be enhanced by second-site mutations to make infections caused by the organism essentially untreatable. Whole genome analyses of the isogenic mutants did not identify significant novel genomic mutations that were shared among 3502APMx or 3502APMx-x mutants. Therefore, genomic mutations alone did not fully explain the MIC patterns observed in the different sets of mutants.

In an attempt to further elucidate the source of these increased MICs we looked at the methylation patterns of 3502 and the isogenic mutants. Initial results from the PacBio base modification detection analyses demonstrated that the 3502 reference, the nine AMP mutants, and a clinical control sample contained several shared m6A motifs and one m4C motif. However, it was also observed that as the ceftriaxone MICs of the isolates increased, the number of modified motifs detected expanded, including some novel motifs that were classified as “unknown” by the software program. This suggests that methylation might play a role in gonococcal antimicrobial resistance.

This study also aimed to characterize the methylation patterns of clinical isolates that are resistant to other antibiotics. We examined 16 clinical gonococcal isolates, including eight ciprofloxacin-resistance isolates with varied MIC values (1-32 µg/mL) and eight azithromycin-susceptible (MIC ≤ 1.0) and resistant (MIC ≥ 2.0) isolates. These results will shed light on the role of epigenetics in antimicrobial resistance.

Utility of Ion Torrent PGM Sequencing in Genomic Characterization of Clostridium botulinum

Friday, 29th May 12.15 - La Fonda Ballroom - Oral

Brian Raphael[1]
[1]National Botulism Laboratory Team, Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention

Botulinum neurotoxins are responsible for the neuromuscular paralysis associated with botulism and are produced by Clostridium botulinum. These strains are genetically diverse and produce specific toxin serotypes (A-G) which are associated with different metabolic/phylogenetic groups of C. botulinum. We utilized the Ion Torrent PGM in various projects aimed at characterizing the phylogeny, subtyping and pathogenesis of this organism.

Core genome phylogeny revealed that Group I C. botulinum and non-toxic C. sporogenes strains could be distinguished into distinct sub-clades. Clade-specific genetic markers were used to examine 24 putative C. sporogenes strains of which 17 strains contained markers specific to C. sporogenes. Genome sequencing of two strains lacking these markers revealed that one strain likely represented C. botulinum that lost a toxin gene while another strain was found to be divergent enough to be considered a separate species.

Reference-free SNP analysis (kSNP) was used to distinguish foodborne-associated strains of C. botulinum type A(B) that were indistinguishable by pulsed-field gel electrophoresis. This analysis demonstrated the ability to differentiate strains from all of the outbreaks examined and a non-outbreak associated strain. Moreover, strains associated with the same outbreak but isolated from different sources clustered together.

Botulism can also occur when the intestines of infants become colonized with C. botulinum. We utilized amplicon sequencing targeting the V3 region of the 16S rRNA gene to analyze the microbial communities in fecal samples from infants. Sequences were analyzed with QIIME which revealed that C. botulinum is found in very low abundance among positive samples and that these samples were enriched for the presence of certain types of proteobacteria.

These findings demonstrate an important role for draft genome sequences in the understanding of the evolutionary dynamics of C. botulinum, distinguishing strains isolated from botulism outbreaks, and in characterizing the fecal microbial communities of infants with botulism.

Keywords: botulism, Ion Torrent PGM, amplicon sequencing, SNP analysis, core genome phylogeny

The findings and conclusions in this presentation are those of the author and do not represent the official position of the Centers for Disease Control and Prevention.

Genetic diversity within the botulinum neurotoxin---producing bacteria and their neurotoxins

Friday, 29th May 12.30 - La Fonda Ballroom - Oral

Karen Hill[1]
[1]**Los Alamos National Laboratory**

The recent availability of multiple Clostridium botulinum genomic sequences has initiated a new genomics era that strengthens our understanding of the bacterial species that express botulinum neurotoxins (BoNTs). Analysis of the genomes has reinforced the historical organism Group I---IV designations and provided evidence that the bont genes can be located within the chromosome, phage or plasmids. The sequences provide the opportunity to examine closely the variation among the toxin genes, the composition and organization of the toxin complex, the regions flanking the toxin complex and the location of the toxin within different bacterial strains. These comparisons provide evidence of horizontal gene transfer and site---specific insertion and recombination events that have contributed to the variation observed among the neurotoxins.

Wrap up of SFAF 2015 and planning for 2016

Friday, 29th May 12.45 - La Fonda Ballroom - Keynote/Plenary

Chris Detter
Los Alamos National Laboratory

Evaluation of Massively Parallel Sequencing Technologies for Expanded DNA Identification Capabilities at the Federal Bureau of Investigation Laboratory

Friday, 29th May 13.30 - La Fonda Ballroom - Forensic

***Jodi Irwin*[1], *Lilliana Moreno*[1], *Anthony Onorato*[1], *Michael Brandhagen*[1], *Thomas Callaghan*[1]**

[1]*Federal Bureau of Investigation*

Though Massively Parallel Sequencing (MPS) has transformed numerous genetic disciplines over the past decade, it is only within the past few years that evaluations of MPS for forensic application have been undertaken in earnest. Given the potential of MPS to not only increase the quantity and discriminatory power of genetic data but also improve the overall throughput of samples through the laboratory, the Federal Bureau of Investigation is exploring the development of MPS assays for possible casework application. Long-term laboratory efforts are directed towards employing MPS as a common platform for testing of all markers of forensic interest. However, near-term efforts are directed specifically towards evaluating the technology for its utility in expanding existing institutional capabilities. Areas of current emphasis are 1) highly challenging samples and the benefits of MPS for improved information recovery 2) mitochondrial DNA typing and the development of entire mitochondrial genome (mtGenome) data in particular and 3) no-subject crime scene samples and the value of ancestry and phenotype markers for developing investigative leads. Commercially available assays designed for forensic application, as well as custom assays for specific nuclear DNA markers and the mtGenome, are currently under evaluation. In addition to its use for these near-term operational applications, MPS is also being used more generally - to better characterize our most challenging samples. With a better understanding of endogenous DNA quantity and quality, forensic examination strategies may be devised that accommodate the challenges of specific case scenarios and therefore broaden the range of sample type and quality from which probative data can be recovered. Here, we present an overview of our general efforts.

Validation of a Targeted Next Generation Sequencing Solution for Forensic Genomics

Friday, 29th May 13.45 - La Fonda Ballroom - Forensic

Ernie Guzman[1], Kathryn Stephens[1], Cydne Holt[1], Joe Valaro[1]
[1]Illumina, Inc

Sequencing (NGS) by Synthesis (SBS) enables the entire human genome to be sequenced in one day. As a simpler yet highly effective alternative, forensic scientists can choose to perform targeted sequencing of PCR products. By sequencing a dense set of forensic loci, casework and database efforts are directed toward the genomic regions that best answer forensic questions, relieving privacy concerns and simplifying analysis. Because it does not depend on allele separation by size, the number of targets interrogated is not limited, allowing a more comprehensive result to be generated.

We describe the development and validation of a targeted amplicon panel and associated bioinformatics tools for forensic genomics that combines and interrogates a core of global short tandem repeat markers used routinely today, along with additional forensic loci that can provide information when standard markers would fail to sufficiently resolve a case. Maximizing the number and types of markers that are analyzed for each sample provides more comprehensive and discriminating information for standard samples, as well as challenging samples that contain low quantities of DNA, degraded and/or inhibited DNA, and complex mixtures. The targeted amplicon panel will enable more complex kinship analysis to be performed, and can also reveal phenotypic and biogeographical ancestry information about a perpetrator to assist with criminal investigations. This capability is expected to dramatically improve the ability to investigate dead end cases, where a suspect reference sample or database hit are not available. We will describe the complete workflow, system, and data analysis tools, and present data from validation and collaborator studies including reproducibility, sensitivity, actual forensic samples, and concordance with standard capillary electrophoresis methods.

Sequence Diversity within Short Tandem Repeat Loci: Applications to Human Identification

Friday, 29th May 14.00 - La Fonda Ballroom - Forensic

***Peter Vallone*[1], Rachel Aponte[1], Michael Coble[1], Kevin Kiesler[1], Katherine Gettings[1]**
[1]National Institute of Standards and Technology*

In this presentation, NGS results from a set of 183 samples from three U.S. groups (Caucasian, African American, and Hispanic) will exemplify the sequence variation that exists in 22 autosomal forensic STR loci. This variation can be categorized as either repeat region sequence motif changes which result in isoalleles (alleles of the same length but different sequence) or flanking region polymorphisms (SNPs or InDels). This experimental sequence data gives an indication of the level of diversity expected in the larger population and will be used to provide examples of how isoalleles and flanking region variants at some loci can improve discrimination and mixture deconvolution in forensic casework.

Human Identification Advances Using Next Generation Sequencing Technologies

Friday, 29th May 14.15 - La Fonda Ballroom - Forensic

***Joseph Chang*[1], *Reina Langit*[1], *Narasimhan Rajagopalan*[1], *Sharon Wootton*[1], *Chien-Wei Chang*[1]**

[1]Thermo Fisher Scientific

Capillary Electrophoresis (CE) is traditionally used to sequence Short Tandem Repeats (STRs) to accurately identify humans. With the ever-increasing human population and the increase of compromised casework samples, there is a need to use alternative technologies to complement incomplete CE STR profiles. Next Generation Sequencing (NGS) technologies have already been highly exploited for accurate, high-throughput whole genome discovery. As such, NGS has the ability to detect sequence variation of the same length between individuals. Furthermore, the high-multiplex capability of a NGS STR assay allows analysts to recover profiles from low input DNA of casework samples. However, NGS should not only target STRs. Rather, by additionally targeting Single Nucleotide Polymorphisms (SNPs) and Mitochondrial DNA, discrimination power exponentially increases. Ancestral-Informative Markers (AIMs) are used as a statement for investigative leads, while Identity-Informative SNPs (iiSNPs) and Mitochondrial DNA are both utilized for high compromised samples. High-Multiplex STR, SNP, and Mitochondrial panels have been developed for the Ion Personal Genome Machine® (PGM™) System. Concordance studies were performed between the NGS STR panel and CE STR kits, while both SNP panels were tested against TaqMan® assays. The Mitochondrial panel was used to recover profiles from challenging hair and bone samples. Complementing CE STR results, NGS panels can be sequenced in parallel to provide more leads with existing forensic databases.

Assessment of Massively Parallel Sequencing chemistries for forensic casework applications

Friday, 29th May 14.30 - La Fonda Ballroom - Forensic

***Lilly Moreno**[1], Michael Brandhagen[1], Anthony Onorato[1], Thomas Callaghan[1], Jodi Irwin[1]*

[1]Federal Bureau of Investigation

Massively Parallel Sequencing (MPS) has emerged as a powerful and cost-effective method for the development of genetic data. For forensic applications, MPS is likely to greatly expand the range of samples from which probative data can be recovered, increase the discriminatory power of DNA information in challenging scenarios and dramatically improve the cost and speed of forensic DNA analysis. As part of an ongoing effort to evaluate MPS technology for potential application to human identification casework, a number of manufacturer produced multiplex assays and software packages were tested on a variety of sample types. The Illumina ForenSeq kit and the Promega PowerSeq Auto/Mito/Y kit were applied to well-characterized control DNAs, as well as a number of high and low-quality samples typical of specimens routinely encountered in forensic casework. Illumina MiSeq-developed MPS data from both kits were processed through various software packages/bioinformatics pipelines and then assessed for accuracy, sensitivity, and information gain, using standard capillary electrophoresis-based Sanger-sequencing and fragment analysis data for comparison. In addition, MiSeq- developed single nucleotide polymorphism (SNP) data were compared to Ion Torrent PGM-developed SNP results to both validate the genotypes and gain further insight into the practical implications of MPS-platform specific features.

A Method for Separating Microbial DNA from Vertebrate DNA

Friday, 29th May 14.45 - La Fonda Ballroom - Forensic

Fiona Stewart[1], Erbay Yigit[1], George Feehery[1], Samuel Oyola[2], Yan Wei Lim[3], David Hernandez[4], Brad Langhorst[1], Joanna Bybee[1], Laurie Mazzola[1], Lynne Apone[1], Christine Chater[1], Pingfang Liu[1], Christine Sumner[1], Donovan Bailey[4], Eileen Dimalanta[1], Theodore Davis[1], Michael Quail[2], Sriharsa Pradhan[1]
[1]New England Biolabs, Inc., [2]Wellcome Trust Sanger Institute, [3]San Diego State University, [4]New Mexico State University

As Next Generation Sequencing (NGS) is becoming more widespread and moving into new application areas, the use of more challenging samples is increasing. Of particular interest are samples of mixed origin in which the DNA of interest is contaminated or overwhelmed by other DNA in the sample.

Microbiome samples fall into this category. Recent discoveries have implicated the human microbiome as playing a role in certain physical conditions and disease states, and these advances have opened up the potential for development of microbiome-based diagnostic and therapeutic tools as well as potential use in forensics applications. However, samples from many sites, including those derived from vertebrate skin, bodily cavities, and body fluids, contain both host and microbial DNA. Since a single human cell contains approximately 1,000 times more DNA than a single bacterial cell, even low-level human cell contamination can substantially complicate the analysis of the microbial content of a sample. In some cases, as little as 1% of sequencing reads may pertain to the microbes of interest and a large percentage of sequencing reads must be discarded, making such experiments impractical. The inverse situation, in which the genomic, nuclear DNA is of interest, but is substantially contaminated by microbial DNA, is similarly challenging.

To address these issues, we developed a method to separate microbial DNA from vertebrate genomic DNA. This method exploits the difference in CpG methylation between vertebrate genomic DNA and microbial genomic DNA. The methyl-CpG binding domain (MBD) binds specifically to vertebrate DNA, which is highly CpG methylated. MBD does not bind to microbial DNA, which has little to no CpG methylation. This differential binding enables separation and capture of the two types of DNA, either of which can be used in subsequent studies. Importantly, microbial diversity and relative abundance is maintained after enrichment.

Here we show that this simple magnetic bead-based method can be used to separate vertebrate DNA from bacterial and protistan DNA in many different sample types in which the vertebrate host DNA dominates, including human saliva, human blood, a mock malaria-infected blood sample and human cystic fibrosis sputum. Sequencing of the microbial enriched fraction, on multiple next generation sequencing platforms, showed a 50-fold reduction in reads aligning to the human genome and a 10- fold increase in reads aligning to microbial genomes. Conversely, sequencing of the captured host fraction showed that the vast majority (>99%) of the mapped reads aligned to the human reference genome.

Beyond microbial DNA analysis, separation on the basis of differential methylation also enables isolation of various organelles' DNA. For example, while human and plant genomic DNA is bound readily by MBD, mitochondrial and chloroplast DNA is not, thereby allowing for enrichment of these organelles from a variety of samples. Differential separation of DNA populations based on methylation status opens up many opportunities in diagnostics, therapeutics and forensic science.

Next-generation Sequencing and Custom Software Analysis of mtDNA Mixtures

Friday, 29th May 15.15 - La Fonda Ballroom - Forensic

Cassandra Calloway[1], Hanna Kim[1], Rachel Gordon[1], Daniela Cuenca[2], Samuel Vohr[3], Richard Green[4], Richard Edward Green[4], George Sensabaugh[5], Henry Erlich[1]
[1]Children's Hospital Oakland & Research Center, [2]university of california, Davis, [3]universiy of California, Santa Cruz, [4]University of California Santa Cruz, [5]University of California Berkeley

Next-generation sequencing technologies have revolutionized the field of genetics and have the potential to make a significant impact to the DNA forensics field. Specifically, the clonal sequencing aspect of these technologies allows for sequencing individual DNA fragments which can be applied to analysis of DNA mixtures often encountered in forensic casework. Current methods which use Sanger sequencing for analysis of mtDNA do not allow for interpretation of mixtures since the peak heights do not accurately represent the individual sequence components of a mixture. Additionally, Sanger sequencing is also limited in detection of minor components of a mixture (10-15%). We present here results analyzed using custom software demonstrating the application of two NGS mtDNA enrichment assays for analysis of mixtures.

A duplex PCR assay targeting the mtDNA hypervariable regions I and II (HVI/HVII) was developed using eight sets of 454 MID tagged fusion primers in a combinatorial approach for deep sequencing 64 samples in parallel on a 454 GS Jr. This assay was shown to be highly sensitive for sequencing limited DNA amounts (~100 mtDNA copies) and detecting mixtures with low level variants (~1%) as well as heteroplasmy. In addition, "complex" mixtures (≥3 contributors) were successfully sequenced and analyzed using custom software. Currently, commercial software is limited for the analysis of mixtures as these softwares only report the frequency of variants detected and not the frequency of the distinct sequence haplotypes. We have developed and tested an algorithm, 'hap-summary.pl', which reports the frequency of the detected sequence haplotypes allowing for analysis of complex mixtures.

We have also developed a solution phase sequence capture and NGS assay for targeted enrichment and deep sequencing of the entire mitochondrial genome for increased discrimination power using Illumina and 454 NGS technologies. Customized SoftGenetics NextGENe software with added features for alignment of the circular mitochondrial genome and mutation reporting following the forensic mtDNA SWGDAM guidelines was used for the alignment and NGS analysis. Using the Sequence Capture NGS assay, 100% sequence coverage of the mitochondrial genome was achieved with an ~95% on target rate using the an optimized library preparation method and Illumina MiSeq compared to ~80% using the 454 GS JR. The depth of coverage across the mitochondrial genome was also more balanced using the Illumina MiSeq compared to the 454 GS Jr.

Evaluation of Concordance and Low Level Variant Detection for Forensic-Quality High-Throughput Sequencing of the Full mtGenome using the Illumina MiSeq

Friday, 29th May 15.30 - La Fonda Ballroom - Forensic

***Michelle Peck*[1], *Michael Brandhagen*[2], *Toni Diegoli*[3], *Jodi Irwin*[2], *Charla Marshall*[1], *Kimberly Sturk-Andreaggi*[1]**

[1]*Armed Forces DNA Identification Laboratory/American Registry of Pathology*, [2]*Federal Bureau of Investigation*, [3]*Armed Forces DNA Identification Laboratory/American Registry of Pathology; Current: Battelle*

As next generation sequencing (NGS) is evaluated for mitochondrial DNA (mtDNA) analysis in forensic laboratories, it is essential to demonstrate the accuracy of this technology. To this end, a concordance study between Sanger-type sequencing (STS) and a NGS methodology was carried out with 90 high-quality serum samples previously processed with STS methods according to forensic standards. The 90 samples, along with six appropriate controls, were processed in a high-throughput (HTP) manner using long range target enrichment of the full mtDNA genome (mtGenome), Nextera® XT library preparation, and sequencing on the Illumina MiSeq. Two libraries were generated at two different laboratories from the same amplicon product, which allowed for comparison of library normalization strategies (quantification-based versus bead-based) and sequencing data from two MiSeq instruments. Analysis was carried out on the CLC Genomics Workbench using a customized workflow, and each sequencing run was independently analyzed by two scientists using different software versions. To interpret the NGS data, an initial 5% low level variant detection threshold was utilized and known regions of length heteroplasmy were ignored due to alignment issues. Comparison of the NGS variant tables and the STS profiles demonstrated 99.9994% concordance with only 19 discordant calls at over 2,900,000 positions analyzed. Of these discrepancies, six point heteroplasmies (PHPs) that were not observed in the STS data were detected with the NGS data because of increased sensitivity. Other discordant calls were either due to minor alignment issues or low coverage. Moreover, two samples were identified as mixtures at 1:20 and 1:50 ratios. The average coverage was approximately 2000X for both libraries when multiplexing 96 samples, though the bead-based library showed more variability in average coverage per sample. While the initial 5% threshold resulted in reliable mtDNA profiles, additional analysis will investigate low level variant detection down to a 1% threshold to give better insight into the level of background noise and potential strategies for distinguishing it from true variants. This data will assist in determining if at this level of multiplexing the same accuracy in profiles can be consistently obtained at a lower detection threshold. Although further refinement of analysis and interpretation guidelines is necessary, this concordance study demonstrated a robust and accurate strategy for sequencing the full mtGenome in a HTP manner and allowed for increased detection of low level variants.

Disclaimer: The opinions or assertions presented hereafter are the private views of the authors and should not be construed as official or as reflecting the views of the Department of Defense, its branches, the U.S. Army Medical Research and Materiel Command, the Armed Forces Medical Examiner System, the Federal Bureau of Investigation or the U.S. Government.

Strengths and Limitations of NGS for Forensic DNA Analysis

Friday, 29th May 15.45 - La Fonda Ballroom - Forensic

Jaynish Patel[1], Spencer Hermanson[1], Douglas Storts[1]
[1]Promega Corporation

The prototype PowerSeq™ Systems include primers and amplification master mix for sequencing autosomal short tandem repeats (STRs), Y-chromosome STRs, the mitochondrial DNA control region, and various combinations of the three on an Illumina MiSeq® System. The selected STR loci are the same as used in the commercial PowerPlex® Fusion and PowerPlex® Y23 Systems that are routinely used for capillary electrophoresis-based forensic DNA analysis. We will present data demonstrating performance of the PowerSeq™ Systems, outlining both the strengths and limitations of current NGS technologies for routine forensic analysis.

A Universal Microbial Clock for Estimating the Postmortem Interval

Friday, 29th May 16.00 - La Fonda Ballroom - Forensic

Jessica Metcalf[1], Zhenjiang Xu[2], Will van Treuren[3], Embriette Hyde[2], Daniel Haarmann[4], Amnon Amir[2], Sophie Weiss[1], Se Jin Song[1], Gail Ackermann[2], Gregory Humphrey[2], David Carter[5], Aaron Lynne[4], Sibyl Bucheli[4], Rob Knight[2]

[1]University of Colorado, [2]University of California San Diego, [3]Stanford University, [4]Sam Houston State University, [5]Chaminade University Honolulu

Establishing the time since death and locating clandestine graves are crucial goals in many forensic investigations, but can often be a challenge because our understanding of human decomposition is very limited. Forensic science currently does not fully leverage decomposition processes as physical evidence; microbial communities in particular are not regularly used as a forensic tool despite their ubiquity and crucial role in decomposition. To explore the utility of microbial communities as a forensic tool, we performed a complementary set of experiments using a mouse model system under controlled laboratory settings and donor human subjects in outdoor settings simulating potentially realistic death scene scenarios. Using amplicon-based, high-throughput marker gene (16S rRNA, 18S rRNA, and ITS) sequencing, we characterized the full microbial diversity of skin and gravesoils during decomposition. We show that bacterial and microbial eukaryotic communities can be used to estimate the postmortem interval (PMI) in both a laboratory-based mouse model system as well as in a field-based human cadaver system. Using random forests regression models to estimate PMI prediction error, we achieved error rates ~5 days over our ~70 day experiments, and as little as ~2-3 days in the first 30 days of each experiment. These errors are similar to or better than estimates based on forensic entomology tools. Furthermore, we used microbial data generated from human cadavers decomposing in winter to train our regression model and accurately predict PMI for human cadavers decomposing in spring. By comparing microbial decomposer communities across experiments, we discovered that some decomposers are universal across host and soil type; indeed, in our mouse model experiment, soil type (desert, shortgrass prairie, or forest) did not influence the PMI model error rates. Additionally, we demonstrate that bodies decomposing on soil modify the soil microbial community substantially, allowing for the detection of a decomposition event using the soil microbial community for up to 30 days after the body has been moved. Together, these results suggest that high-throughput sequencing of microbial DNA associated with decomposition can be a powerful forensic science tool, particularly in the context of estimating PMI or locating clandestine graves.

Microbial Forensics of select agents from trace environmental or clinical samples: making the case for targeted sequencing

Friday, 29th May 16.15 - La Fonda Ballroom - Forensic

Tom Slezak[1]
[1]***Lawrence Livermore National Lab***

Making rapid and accurate phylogenetic assignments of select agents from trace environmental or clinical samples is a currently un-met need. Either the agent needs to be cultured (if viable agent is still present) or one must perform 4-7 days of expensive deep WGS metagenomics sequencing and hope to get sufficient reads from the select agent to make confident phylogenetic assignment. The alternative that can be considered is to perform targeted sequencing using commercially available systems from several sequencing platform vendors that permit custom focusing on the organism, gene, and/or SNP regions that will provide you with maximally-usable information at whatever resolution(s) you need for your purpose. We will discuss our experience with such systems and why we think that microbial forensics and clinical diagnostics are 2 uses of NGS that should be moving strongly in this direction.

Bioforensic Metagenomics at the National Bioforensic Analysis Center

Friday, 29th May 16.30 - La Fonda Ballroom - Forensic

M. J. Rosovitz[1]
[1]**National Biodefense Analysis and Countermeasures Center (NBACC)**

The ultimate goal of metagenomic analysis is to analyze any type of sample to identify and quantify all of its biological components. While on the surface this might appear straightforward, each step in the process from nucleic acid extraction through bioinformatic analysis presents potential pitfalls that can bias analysis results. In the laboratory, the challenges revolve around generating sequence data that represents everything in the sample without introducing contaminating sequence from the laboratory, workers, and reagents. On the bioinformatics side, the difficulties pertain to using fragmented sequence data to fully reconstruct a biological sample. Additionally, different bioforensic applications require different metagenomic methods, necessitating a flexible and multifaceted approach. Here we present analyses of simple and complex samples at NBFAC, using examples to discuss how our bioforensic metagenomics approach is currently applied for each method. Ongoing challenges for the field are highlighted with respect to metagenomics, in general, as well as its application to bioforensics.

Round Table Discussion of Forensics Applications for NGS Technologies

Friday, 29th May 16.45 - La Fonda Ballroom

***Robert Bull*, Chair**
Department of Homeland Security

Attendee Listing

First Name	Last Name	Company / Organization	Email Address
Audrey	Abrams McLean	Centers for Disease Control and Prevention	aabramsmclean@cdc.gov
Omayma	Al-Awar	Illumina	oalawar@illumina.com
Johar	Ali	Alvi Armani	ali.johar@gmail.com
Ann	Allison	Illumina	aallison@illumina.com
Michael	Alonge	Driscoll Strawberry Associates	michael.alonge@driscolls.com
Murtada	Alsaadi	University of New Mexico	malsaadi@salud.unm.edu
Samar	Alshorman	Jordan University of Science and Technology	smalshorman@just.edu.jo
Mohammad Ruhul	Amin	Stony Brook University	mohammad.r.amin@stonybrook.edu
Jason	Anderson	Liberty Biosecurity	jason@libertybiosecurity.com
Joe	Anderson	Defense Threat Reduction Agency	joseph.anderson@dtra.mil
Taylor	Appleberry	PerkinElmer, Inc.	taylor.appleberry@perkinelmer.com
Arthur	Armijo	University of New Mexico	arthur11@unm.edu
Jason	Aulds	US Department of Defense	jaulds@nmci.detrick.army.mil
Robert	Baker	Texas Tech University	robert.baker@ttu.edu
jack	ballantyne	University of Central Florida	jack.ballantyne@ucf.edu
Jim	Bartholomew	JCB Scientific Consulting	jcbarthogone@yahoo.com
Dhwani	Batra	Chanega CGC, Representing CDC	bun3@cdc.gov
Joseph	Baugher	US Food and Drug Administration CFSAN	joseph.baugher@fda.hhs.gov
Callum	Bell	National Center for Genome Resources (NCGR)	cjb@ncgr.org
Nicolas	Berthet	CIRMF	nicolas.berthet@ird.fr
Jasbir	Bhangoo	Driscoll Strawberry Associates	jasbir.bhangoo@driscolls.com
Jonathan	Bingham	Google	binghamj@google.com
Sven	Bocklandt	BioNano Genomics	SBocklandt@bionanogenomics.com
Eric	Boerwinkle	University of Texas	Eric.Boerwinkle@uth.tmc.edu
Joseph	Bogan	MRIGlobal	jbogan@mriglobal.org
Cecilie	Boysen	Qiagen Aarhus	cecilie.boysen@qiagen.com
Catherine	Branda	Sandia National Laboratories	cbranda@sandia.gov
Gavin	Braunstein	Defense Threat Reduction Agency	gavin.m.braunstein.civ@mail.mil
Raquel	Bromberg	University of Texas Southwestern Medical Center	Raquel.Bromberg@utsouthwestern.edu
David	Bruce	Los Alamos National Laboratory	dbruce@lanl.gov
Lijing	Bu	University of New Mexico	lijing@unm.edu
Christian	Buhay	Baylor College of Medicine / HGSC	cbuhay@bcm.edu
Robert	Bull	Federal Bureau of Investigation Laboratory	robert.bull@associates.hq.dhs.gov
Scott	Burns	Centers for Disease Control and Prevention / BAH	yqd0@cdc.gov
Carlos	Bustamante	Liberty Biosecurity	carlos@libertybiosecurity.com
Jason	Byars	University of New Mexico	jbyars@unm.edu
Timothy	Byaruhanga	Uganda Virus Research Institute (UVRI)	tssekandi@gmail.com
Bernarda	Calla	US Department of Agriculture ARS	bernarda.calla@ars.usda.gov
Thomas	Callaghan	Federal Bureau of Investigation	thomas.callaghan@ic.fbi.gov
Cassandra	Calloway	Children's Hospital & Research Center at Oakland	scalloway@chori.org
Heather	Carleton	Centers for Disease Control and Prevention	hcarleton@cdc.gov
Michael	Cassler	MRIGlobal	mcassler@mriglobal.org
Patrick	Chain	Los Alamos National Laboratory	pchain@lanl.gov
Joseph	Chang	Thermo Fisher Scientific	joseph.chang@thermofisher.com
Gvantsa	Chanturia	National Center for Disease Control and Public Health (Georgia)	gvantsa.chanturia@ncdc.ge
Sihong	Chen	Thermo Fisher Scientific	sihong.chen@lifetech.com
Olga	Chertkov	Los Alamos National Laboratory	ochrtkv@lanl.gov
Jason	Chin	Pacific Biosciences	jchin@pacificbiosciences.com
William	Chow	Wellcome Trust Sanger Institute	wc2@sanger.ac.uk
Wai Kwan	Chung	BNBI	chungw@nbacc.net
Christohper	Citraro	Thermo Fisher Scientific	chris.citraro@lifetech.com

Alicia	Clum	Lawrence Berkeley National Laboratory JGI	aclum@lbl.gov
Rebecca	Colman	Translational Genomics Research Institute	rcolman@tgen.org
Sean	Conlan	NHGRI	conlans@mail.nih.gov
Deidra	Conner	Chemagen - A PerkinElmer Company	deidra.conner@perkinelmer.com
Helen	Cui	Los Alamos National Laboratory	hhcui@lanl.gov
Hajnalka	Daligault	Los Alamos National Laboratory	hajkis@lanl.gov
Ashlynn	Daughton	Los Alamos National Laboratory	adaughton@lanl.gov
Karen	Davenport	Los Alamos National Lab	kwdavenport@lanl.gov
Matthew	Davenport	Department of Homeland Security / S&T Directorate	matthew.davenport@hq.dhs.gov
Josie	Delisle	Translational Genomics Research Institute	jdelisle@tgen.org
Chris	Detter	LANL / DTRA	cdetter@lanl.gov
Armand	Dichosa	Los Alamos National Laboratory	armand@lanl.gov
Tamar	Dickerson	MRIGlobal	tdickerson@mriglobal.org
Todd	Dickinson	Dovetail Genomics, LLC	todd@dovetail-genomics.com
Toni	Diegoli	Battelle	diegoli@battelle.org
Darrell	Dinwiddie	University of New Mexico	dldinwiddie@salud.unm.edu
Norman	Doggett	Los Alamos National Laboratory	doggett@lanl.gov
Matt	Dunn	Wellcome Trust Sanger Institute	md3@sanger.ac.uk
Zeljko	Dzakula	BioNano Genomics	zdzakula@bionanogenomics.com
donald	eby	Booz Allen Hamilton	eby_donald@bah.com
Adam	English	Human Genome Sequencing Center	english@bcm.edu
Tracy	Erkkila	Los Alamos National Laboratory	terkkila@lanl.gov
Henry	Erlich	Children's Hospital & Research Center at Oakland	herlich@chori.org
Seth	Faith	North Carolina State University	safaith@ncsu.edu
Lawrence S.	Fakoli III	Liberian Institute for Biomedical Research	lawfako2008@yahoo.com
Jason	Farlow	Farlow Scientific Consulting LLC	farlowscience@gmail.com
Bethaney	Fehrenkamp	University of New Mexico	bfehrenk@unm.edu
shihai	feng	Los Alamos National Laboratory	sfeng@lanl.gov
Kevin	Fengler	Dupont Pioneer	kevin.a.fengler@pioneer.com
Haley	Fiske	Illumina	hfiske@illumina.com
Michael	FitzGerald	Broad Institute of MIT and Harvard	fitz@broadinstitute.org
Michael	Franklin	Montana State University	franklin@montana.edu
Nicole	FRAZIER	DFSC - USACIL	nicole.r.frazier.civ@mail.mil
Robert	Fulton	Washington University in St. Louis	bfulton@genome.wustl.edu
James	Gale	Tricore Reference Lab	James.Gale@tricore.org
Scott	Geib	US Department of Agriculture ARS	scott.geib@ars.usda.gov
John	Gillece	Translational Genomics Research Institute North	jgillece@tgen.org
Lori	Gladney	Centers for Disease Control and Prevention	hze1@cdc.gov
Cheryl	Gleasner	Los Alamos National Laboratory	cdgle@lanl.gov
Veena	Gnanakkan	USG	fp.veena@gmail.com
Darren	Grafham	Sheffield Children's NHS foundation trust	darren.grafham@sch.nhs.uk
Richard	Guerrieri	Battelle	guerrierir@battelle.org
Stephanie	Guida	National Center for Genome Resources (NCGR)	sguida@ncgr.org
Jeffrey	Gunter	Department of Defense	jngunter@gmail.com
Jon	Hagopian	Advanced Analytical	jhagopian@aati-us.com
David	Hall	Defense Threat Reduction Agency	david.j.hall24.civ@mail.mil
Timothy	Hamp	PerkinElmer, Inc.	Timothy.Hamp@perkinelmer.com
William	Hansen	Thermo Fisher Scientific	william.hansen@thermofisher.com
John	Hanson	Research and Testing Laboratory	j.delton.hanson@researchandtesting.com
Tim	Harkins	Swift Biosciences	harkins@swiftbiosci.com
Kevin	Harrod	University of Alabama Birmingham	kharrod@uab.edu
Emily	Hatas	Pacific Biosciences	ehatas@pacb.com
Andrew	Hatch	Los Alamos National Laboratory	ahatch@lanl.gov
John	Havens	Integrated DNA Technologies	jhavens@idtdna.com
Crystal	Hepp	Center for Microbial Genetics and Genomics	crystal.hepp@nau.edu
maria	hernandez	Department of Defense	mariaolgah@gmail.com
Karen	Hill	Los Alamos National Laboratory	khill@lanl.gov
David	Hirschberg	University of Washington	dlhirschberg@gmail.com
Charles	Hong	Defense Threat Reduction Agency	charles.hong@dtra.mil

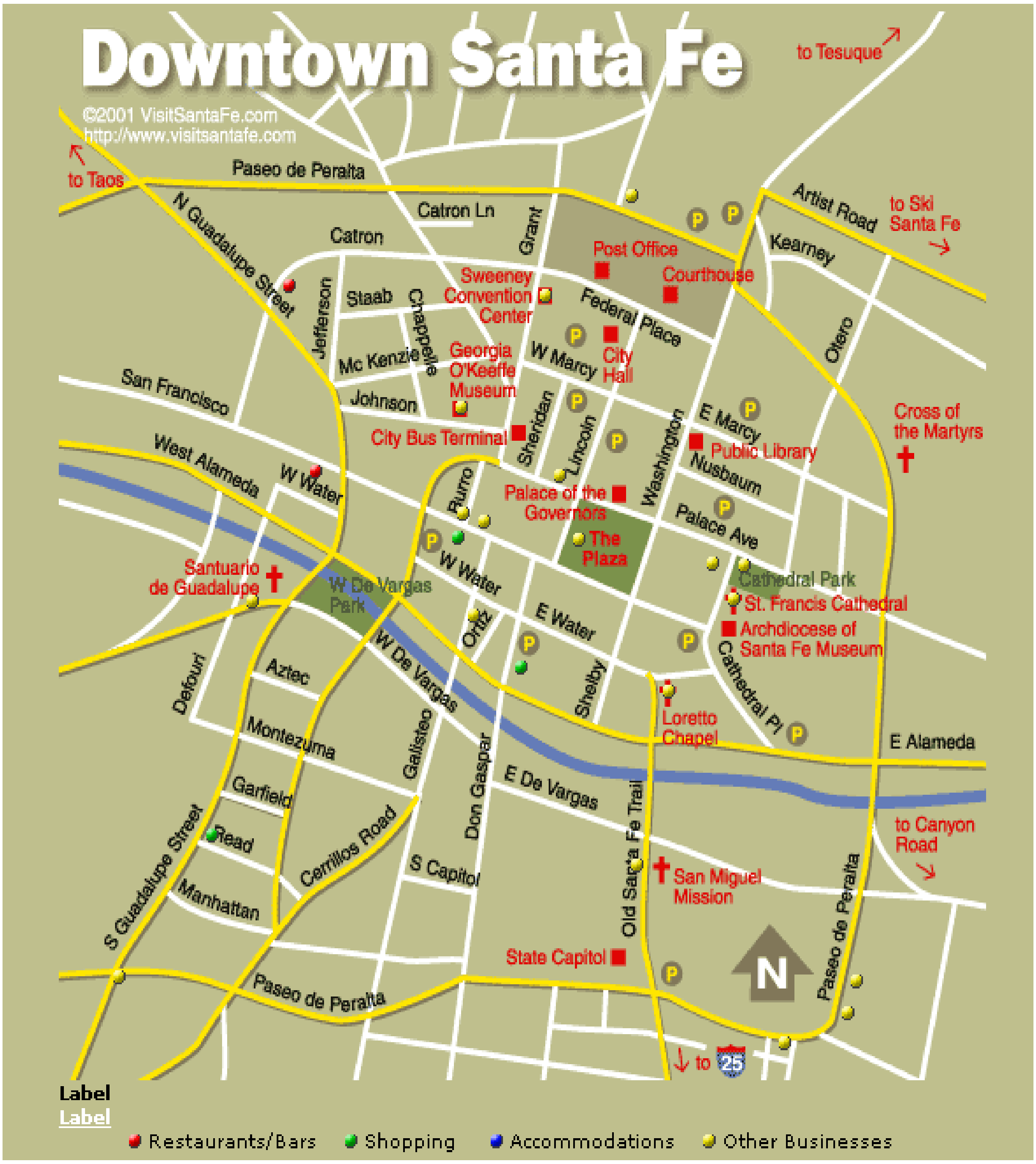
Kelly	Hoon	Illumina	khoon@illumina.com
Chris	Hopkins	Centers for Disease Control and Prevention / BAH	vqd8@cdc.gov
Andrew	Huang	Centers for Disease Control and Prevention	wwm8@cdc.gov
Corey	Hudson	Sandia National Laboratories	cmhudso@sandia.gov
Bill	Huff	Defense Threat Reduction Agency	william.b.huff2.civ@mail.mil
Mariela	Humphrey	Thermo Fisher Scientific	mariela.humphrey@thermofisher.com
Sung	Im	Centers for Disease Control and Prevention	wla9@cdc.gov
Jodi	Irwin	Federal Bureau of Investigation	jodi.irwin@ic.fbi.gov
Rashedul	Islam	University of British Columbia	rashedul.gen@gmail.com
Jonathan	Jacobs	MRIGlobal	jjacobs@mriglobal.org
David	Jaffe	10X Genomics	jaffe@broadinstitute.org
Xiaoben	Jiang	University of New Mexico	sdpapet@unm.edu
Elizabeth	Johnson	U.S. Army Crime Lab	elizabeth.d.johnson38.civ@mail.mil
Shannon	Johnson	Los Alamos National Laboratory	shannonj@lanl.gov
Tracey	Johnson	Tetracore Inc.	tjohnson@tetracore.com
John	Julias	Department of Homeland Security / S&T Directorate	John.Julias@hq.dhs.gov
Lee	Katz	Centers for Disease Control and Prevention EDLB	gzu2@cdc.gov
John	Kayiwa	Uganda Virus Research Institute (UVRI)	jkayiwa@uvri.go.ug
Jonathan	Kayondo	Uganda Virus Research Institute (UVRI)	jkayondo@gmail.com
Paul	Keim	Northern Arizona University	paul.keim@nau.edu
Jon	Kennedy	Noblis	jon.kennedy@noblis.org
Abid	Khan	COMSATS, Abbottabad,	abidkhanuop@gmail.com
Ekaterine	khmaladze	National Center for Disease Control and Public Health (Georgia)	e.khmaladze@ncdc.ge
Seongwon	Kim	Naval Research Laboratory	kittim1000@gmail.com
Luke	Kingry	Centers for Disease Control and Prevention	vtx8@cdc.gov
Kristen	Knipe	Centers for Disease Control and Prevention	wgg9@cdc.gov
Gerwald	Koehler	Oklahoma State University	gerwald.kohler@okstate.edu
Lars	Koenig	Research and Testing Laboratory	lars.koenig@researchandtesting.com
Frank	Kolakowski	Tetracore, Inc.	fkolakowski@tetracore.com
Ara	Kooser	University of New Mexico	ghashsnaga@gmail.com
Anton	Korobeynikov	Saint Petersburg State University	anton@korobeynikov.info
Nato	Kotaria	National Center for Disease Control and Public Health (Georgia)	n_kotaria@yahoo.com
Adam	Kotorashvili	National Center for Disease Control and Public Health (Georgia)	adam.kotorashvili@gmail.com
Paul	Kotturi	Pacific Biosciences	pkotturi@pacb.com
Jochen	Kumm	Pinpoint Science	jochen.kumm@gmail.com
Yuliya	Kunde	Los Alamos National Laboratory	y.a.kunde@lanl.gov
ingrid	labouba	CIRMF (Franceville - Gabon)	ilabouba@gmail.com
Yvonne	Lai	University of California San Francisco	YukYin.Lai@ucsf.edu
Abizar	LAKDAWALLA	Thermo Fisher Scientific	abizar.lakdawalla@thermofisher.com
Ka-Kit	Lam	University of California Berkeley	kakitone@gmail.com
Alla	Lapidus	St.Petersburg State University	yevalmi@gmail.com
Ana	Lauer	Centers for Disease Control and Prevention	YBP6@cdc.gov
Jason	LeBlanc	DFSC	jason.j.leblanc9.ctr@mail.mil
Jeremy	Ledermann	Centers for Disease Control and Prevention	jledermann@cdc.gov
Darrin	Lemmer	Translational Genomics Research Institute North	dlemmer@tgen.org
Poe	Li	Los Alamos National Laboratory	po-e@lanl.gov
Ashley	Linares	Novozymes	akli@novozymes.com
Tina	Lindsay	Washington University in St. Louis	tgraves@genome.wustl.edu
Pingfang	Liu	New England Biolabs	liu@neb.com
Chienchi	Lo	Los Alamos National Laboratory	chienchi@lanl.gov
Chad	Locklear	Integrated DNA Technologies	CLOCKLEAR@IDTDNA.COM
Lijun	Lu	University of New Mexico	lijun80@unm.edu
Duncan	MacCannell	Centers for Disease Control and Prevention	fms2@cdc.gov
Ann	Machablishvili	National Center for Disease Control and Public Health	a_machablishvili@hotmail.com

		(Georgia)	
Mohammed-Amin	Madoui	CEA-Genoscope	amadoui@genoscope.cns.fr
Amjad	Mahasneh	Jordan University of Science and Technology	amjada@just.edu.jo
Cheriece	Margiotta	Los Alamos National Laboratory	cmargiotta@lanl.gov
Arne	Materna	Qiagen Aarhus	Arne.Materna@qiagen.com
Jeff	Maughan	Brigham Young University	Jeff_Maughan@byu.edu
Franklin	Mayanja	Ministry of Agriculture Animal Industry and Fisheries	mayanjaf@gmail.com
Carl	Mayers	Defence Science and Technology Laboratory (DSTL)	cnmayers@dstl.gov.uk
Kristen	McCabe	Los Alamos National Laboratory	kjmccab@lanl.gov
Mitch	McGrath	US Department of Agriculture ARS	mitchmcg@msu.edu
Kim	McMurry	Los Alamos National Laboratory	kmcmurry@lanl.gov
David	Mead	Lucigen	dmead@lucigen.com
Kelly	Meiklejohn	Federal Bureau of Investigation Laboratory	kelly.meiklejohn@ic.fbi.gov
Charles	Melancon	University of New Mexico	cemelanc@unm.edu
Qingchang	Meng	Baylor College of Medicine / HGSC	qmeng@bcm.edu
Amanda	Mercer	Los Alamos National Laboratory	a.n.mercer@me.com
Anthony	Messer	Defence Science and Technology Laboratory (DSTL)	messer23@hotmail.com
Ginger	Metcalf	Baylor College of Medicine / HGSC	metcalf@bcm.edu
Jessica	Metcalf	University of Colorado	jessicalmetcalf@gmail.com
Rob	Miller	University of New Mexico	rdmiller@unm.edu
Timothy	Minogue	USAMRIID	timothy.d.minogue.civ@mail.mil
Samuel	Minot	One Codex	sam@onecodex.com
Yimam Getaneh	Misganie	Ethiopian Public Health Institute	yimamgetaneh@gmail.com
Mari	Miyamoto	Qiagen Aarhus	mari.miyamoto@qiagen.com
Lilliana	Moreno	Federal Bureau of Investigation	lilliana.moreno@ic.fbi.gov
Joann	Mudge	National Center for Genome Resources (NCGR)	jm@ncgr.org
Shwetha	Murali	Baylor College of Medicine / HGSC	ShwethaCanchi.Murali@bcm.edu
Mari	Murtskhvaladze	National Center for Disease Control and Public Health (Georgia)	dna_lab@iliauni.edu.ge
Beth	Mutai	USAMRU-K / KEMRI	beth.mutai@usamru-k.org
Donna	Muzny	Baylor College of Medicine / HGSC	donnam@bcm.edu
Madhugiri	Nageswara-Rao	New Mexico State University	mnrao@nmsu.edu
Gladys	Nakanjako	Ministry of Agriculture Animal Industry and Fisheries	gladyskiggundu@yahoo.com
Vishal	Nayak	SRA International Inc	vishal_nayak@sra.com
Beth	Nelson	Novozymes	bane@novozymes.com
Scott	Ness	University of New Mexico	sness@salud.unm.edu
Judy	Ney	Kapa Biosystems	judy.ney@kapabiosystems.com
Minh	Nguyen	National Institute of Justice	Minh.Nguyen@usdoj.gov
Andriniaina Andy	Nkili Meyong	CIRMF (Franceville - Gabon)	andynkili@gmail.com
Diana	Northup	University of New Mexico	dnorthup@unm.edu
Kristen	O'Connor	Defense Threat Reduction Agency	Kristen.OConnor@dtra.mil
Juliette	Ohan	Los Alamos National Laboratory	johan@lanl.gov
kazufusa	okamoto	DFSC	kazufusa.c.okamoto.ctr@mail.mil
Jose	Olivares	Los Alamos National Laboratory	olivares@lanl.gov
John	Oliver	Nabsys Inc.	oliver@nabsys.com
Christian	Olsen	Biomatters, Inc.	christian@biomatters.com
LUICER	OLUBAYO	USAMRU-K	Luiser.ingasia@usamru-k.org
Zbyszek	Otwinowski	University of Texas Southwestern Medical Center	zbyszek@work.swmed.edu
Oliver	Oviedo	Los Alamos National Laboratory	oviedo@lanl.gov
David	Owuor	Centers for Disease Control and Prevention Kenya	COwuor@kemricdc.org
Clint	Paden	Centers For Disease Control and Prevention	cpaden@cdc.gov
Justin	Page	Brigham Young University	jtpage68@gmail.com
Andy Wing Chun	Pang	BioNano Genomics	apang@bionanogenomics.com
Beverly	Parson-Quintana	Los Alamos National Laboratory	bapq@lanl.gov
Jennifer	Pavlica	Kapa Biosystems	jennifer.pavlica@kapabiosystems.c

			om
Justin	Payne	FDA Center for Food Safety and Applied Nutrition	Justin.Payne@fda.hhs.gov
Michelle	Peck	Armed Forces Medical Examiner System	michelle.a.peck3.ctr@mail.mil
Andreas Sand	Pedersen	Qiagen Aarhus	Andreas.Pedersen@qiagen.com
Pavel	Pevzner	University of California San Diego	ppevzner@cs.ucsd.edu
Gavin	Pickett	University of New Mexico	ggpickett@salud.unm.edu
Thomas	Piggot	Defence Science and Technology Laboratory (DSTL)	tompiggot@hotmail.com
Martin	Pippel	Heidelberg Institute for Theoretical Studies	martin.pippel@h-its.org
Roy N.	Platt	Texas Tech University	neal.platt@gmail.com
Sandra	Porter	Digital World Biology	digitalbio@gmail.com
Jamie	Posey	Centers for Disease Control and Prevention	jposey@cdc.gov
Daniela	Puiu	Johns Hopkins University	dpuiu@jhu.edu
Nicholas	Putnam	Dovetail Genomics, LLC	nik@dovetail-genomics.com
Kashef	Qaadri	Biomatters, Inc.	kashef@biomatters.com
Areej	Quran	Jordan University of Science and Technology	areejalquran@yahoo.com
Mpho	Rakgotho	Zoonotic Research Unit Pretoria	mpho.rakgotho@up.ac.za
Thiru	Ramaraj	National Center for Genome Resources (NCGR)	tr@ncgr.org
Teri	Rambo Mueller	Roche	teri.mueller@roche.com
David	Rank	Pacific Biosciences	drank@pacb.com
Brian	Raphael	Centers for Disease Control and Prevention	BRaphael@cdc.gov
David	Ray	Texas Tech University	david.4.ray@gmail.com
Cassie	Redden	Naval Medical Research Center	cassie.l.redden.ctr@mail.mil
Eric	Rees	RTL Genomics	eric.rees@researchandtesting.com
Brandon	Rice	Dovetail Genomics, LLC	brandon@dovetail-genomics.com
James M	Robertson	Federal Bureau of Investigation Laboratory	james.m.robertson@ic.fbi.gov
Daniel	Rokhsar	DOE Joint Genome Institute, UC Berkeley, LBNL	dsrokhsar@gmail.com
MJ	Rosovitz	National Biodefence Analysis and Countermeasures Center (NBACC)	rosovitzmj@nbacc.net
Raul	Ruiz	US Department of Agriculture	raul.a.ruiz@aphis.usda.gov
Jason	Sahl	Translational Genomics Research Institute	jasonsahl@gmail.com
William	Salerno	Baylor College of Medicine / HGSC	ws144320@bcm.edu
Joe	Salvatore	Qiagen Aarhus	joe.salvatore@qiagen.com
Melanie	Sanchez-Dinwiddie	University of New Mexico	melasanc@unm.edu
Rashesh	Sanghvi	Baylor College of Medicine / HGSC	rsanghvi@bcm.edu
Leif	Schauser	Qiagen Aarhus	leif.schauser@qiagen.com
Melissa	Scheible	North Carolina State University	mkscheib@ncsu.edu
Kelly	Schilling	National Center for Genome Resources (NCGR)	kschilling@ncgr.org
Edward	Schulak	Liberty Biosecurity	edward@schulak.com
Edward	Schulak	MetroBiotech, LLC	edward@schulak.com
Niranjan	Shekar	Spiral Genetics	niranjan@spiralgenetics.com
Sarah	Sheldon	Centers for Disease Control and Prevention	hso5@cdc.gov
Mili	Sheth	Centers for Disease Control and Prevention	shethmili@gmail.com
Palak	Sheth	BioNano Genomics	psheth@bionanogenomics.com
Giwon	Shin	Stanford University	gwonshin@stanford.edu
Brian	Shirey	Centers for Disease Control and Prevention	TShirey@cdc.gov
Michael	Shoemaker	USA/SLAA	mvshoemaker@verizon.net
Heike	Sichtig	US Food and Drug Administration	Heike.Sichtig@fda.hhs.gov
Steve	Siembieda	Advanced Analytical	ssiembieda@aati-us.com
Sheina	Sim	University of Hawaii, Manoa	ssim8@hawaii.edu
Gary	Simpson	University of New Mexico	garyl.simpson@comcast.net
David	Sinclair	Harvard Medical School	david_sinclair@hms.harvard.edu
Nick	Sisneros	Pacific Biosciences	nsisneros@pacb.com
Anthony	Smith	National Institute for Communicable Diseases	anthonys@nicd.ac.za
jason	smith	Pacific Biosciences	jrsmith@pacificbiosciences.com
Todd	Smith	Digital World Biology	tsmith423@gmail.com
Shanmuga	Sozhamannan	Critical Reagents Program, MCS, JPEO	Shanmuga.Sozhamannan.ctr@mail.mil
Ganesh	Srinivasamoorthy	SRA International/ Centers for Disease Control	sganesh02@yahoo.com

Shawn	Starkenburg	Los Alamos National Laboratory	shawns@lanl.gov
Scott	Steelman	Broad Institute of MIT and Harvard	steelman@broadinstitute.org
Fiona	Stewart	New England Biolabs	stewart@neb.com
Jennifer	Stone	MRIGlobal	jstone@mriglobal.org
Dylan	Storey	University of California , Davis	dstorey@ucdavis.edu
Doug	Storts	Promega Corporation	doug.storts@promega.com
Sowmya	Subramanian	New Mexico Consortium	subraman@newmexicoconsortium.org
Ric	Sugarek	Integrated DNA Technologies	rsugarek@idtdna.com
Yongming	Sun	Thermo Fisher Scientific	yongming.sun@thermofisher.com
Anitha	Sundararajan	National Center for Genome Resources (NCGR)	asundara@ncgr.org
Eldin	Talundzic	Centers for Disease Control and Prevention	etalundzic@cdc.gov
Rob	Tarbox	10X Genomics	rob@10xgenomics.com
Cheryl	Tarr	Centers for Disease Control and Prevention	ctarr@cdc.gov
Karen	Taylor	Noblis	karen.taylor@noblis.org
Ken	Taylor	WaferGen Biosystems	ken.taylor@wafergen.com
Krista	Ternus	Signature Science, LLC	kternus@signaturescience.com
Jason	Tidwell	US Department of Agriculture ARS	jason.tidwell@ars.usda.gov
Masoud	Toloue	Bioo Scientific	mtoloue@biooscientific.com
Rick	Tontarski	Defense Forensic Science Center	richard.e.tontarski.civ@mail.mil
Angie	Trujillo	Centers for Disease Control and Prevention	ATrujillo@cdc.gov
Alex	Tumusiime	Centers for Disease Control and Prevention	hma7@cdc.gov
Steven	Turner	Pacific Biosciences	sturner@pacificbiosciences.com
Eishita	Tyagi	Centers for Disease Control and Prevention / BAH	vjn0@cdc.gov
Joshua	Udall	Brigham Young University	jaudall@gmail.com
Sagar	Utturkar	University of Tennessee	sutturka@utk.edu
Willy	Valdivia	Orion Integrated Biosciences Inc.	willy.valdivia@orionbio.com
Peter	Vallone	National Institute of Standards and Technology	peter.vallone@nist.gov
Eric	Van Gieson	MRIGlobal	ericvangieson74@gmail.com
Eric	VIncent	Promega Corp.	eric.vincent@promega.com
Momo	Vuyisich	Los Alamos National Laboratory	vuyisich@hotmail.com
Bonventure	Wachekone	Centers for Disease Control and Prevention Kenya	xwl2@cdc.gov
Darlene	Wagner	Centers for Disease Control and Prevention	ydn3@cdc.gov
Edward	Wakeland	Department of Immunology, UT Southwestern Med. Ctr.	edward.wakeland@utsouthwestern.edu
Bruce	Walker	Applied Invention	bw@ai.co
Kimberly	Walker	Baylor College of Medicine / HGSC	kw5@bcm.edu
Ron	Walters	Pacific Northwest Nat'l. Laboratory	ron@ron-walters.com
Judson	Ward	Driscoll Strawberry Associates	jud.ward@gmail.com
Ian	Watson	Defense Threat Reduction Agency CBEP	ian.watson@dtra.mil
Michael	Weigand	Centers for Disease Control and Prevention	mweigand@cdc.gov
Ryan	Weil	Lockheed Martin	ryan.weil@gmail.com
Neil	Weisenfeld	Broad Institute of MIT and Harvard	neilw@broadinstitute.org
Jacqueline	Weyer	National Institute for Communicable Diseases	jacquelinew@nicd.ac.za
Jeremy	Wilkinson	Research and Testing Laboratory	Jeremy.wilkinson@researchandtesting.com
Diana	Williams	Defense Forensic Science Center	diana.w.williams8.civ@mail.mil
Kelly	Williams	Sandia National Laboratories	kpwilli@sandia.gov
Aye	Wollam	Washington University in St. Louis	awollam@genome.wustl.edu
Jonathan	Wood	Wellcome Trust Sanger Institute	jmdw@sanger.ac.uk
stephen	wyatt	Advantage Genomics	stephenmwyatt@gmail.com
gary	xie	Los Alamos National Lab	xie@lanl.gov
Lindsey	Yoder	Oklahoma State University	lindsey.yoder@okstate.edu
Sarah	Young	Broad Institute of MIT and Harvard	stowey@broadinstitute.org
Stephen	Young	University of New Mexico	steve.young@tricore.org
Ekaterine	Zhgenti	National Center for Disease Control and Public Health (Georgia)	eka_zh@hotmail.com
David	Zorikov	National Center for Disease Control and Public Health (Georgia)	zorikov@gmail.com

Map of Santa Fe, NM



SFAF 2015 Sponsors

illumina

Roche

Promega

PACIFIC
BIOSCIENCES®

SFAF 2015 Sponsors

SFAF 2015 Sponsors

PerkinElmer®
For the Better

BioNano
GENOMICS

WAFERGEN
BIOSYSTEMS

ADVANCED
ANALYTICAL

10X™
GENOMICS

ThermoFisher
SCIENTIFIC

LABCYTE™

BIO

BIOO SCIENTIFIC
MAXIMIZE SCIENCE FOR LIFE®
BIOO LIFE SCIENCE PRODUCTS

Dovetail
GENOMICS,LLC